N° d'ordre : 3913

THÈSE

Présentée devant

devant l'université de Rennes 1

pour obtenir

le grade de : DOCTEUR DE L'UNIVERSITÉ DE RENNES 1
Mention BIOLOGIE

par

Frédéric MAHÉ

Équipe d'accueil : Mécanismes à l'origine de la biodiversité
UMR–CNRS 6553 ECOBIO
École doctorale : Vie – Agro – Santé
Composante universitaire : SCIENCES DE LA VIE ET DE L'ENVIRONNEMENT

Titre de la thèse :

Phylogénie, éléments transposables et évolution de la taille des génomes chez les lupins

soutenue le 17 décembre 2009 devant la commission d'examen

M. :	Philippe	VANDENKOORNHUYSE	Président
M. & M^{me} :	Abdel-Kader	AÏNOUCHE	co-directeurs
	Marie-Thérèse	MISSET	
M^{mes} :	Marie-Angèle	GRANDBASTIEN	Rapporteurs
	Angélique	D'HONT	
M. :	Jeffrey J.	DOYLE	Examinateur

À mon frère

Remerciements

Cette thèse a été réalisée avec le soutien financier du Conseil régional de Bretagne (allocation de recherche doctorale sur le projet GENADAPT 211-B2-9/ARED) et de l'UMR-CNRS 6553 Ecobio. Qu'ils trouvent ici le témoignage de ma gratitude et l'expression de mes sincères remerciements.

En premier lieu, je voudrais remercier Abdel-Kader Aïnouche et Marie-Thérèse Misset, mes co-encadrants, pour m'avoir accueilli dans l'équipe « Évolution des Génomes et Spéciation » de l'UMR Ecobio et m'avoir guidé tout au long de la réalisation de cette thèse. Mes remerciements vont également à Malika Aïnouche (responsable de l'équipe EGS) pour sa disponibilité et l'aide qu'elle m'a apportée à diverses étapes de mon travail.

Je remercie également Angélique D'Hont (CIRAD, Montpellier), Marie-Angèle Grandbastien (INRA, Versailles), Jeffrey J. Doyle (University of Cornell, USA) et Philippe Vandenkoornhuyse (UMR-CNRS Ecobio, Rennes), pour m'avoir fait l'honneur d'évaluer mon travail de thèse.

Merci également à mon tuteur de thèse Alain Bouchereau (UFR-SVE, Rennes-1), à Karine Alix-Jenczewski (AgroParisTech) et à Malika Aïnouche pour leurs critiques et conseils lors de la tenue de mes comités de thèse.

Je souhaite également remercier ici tous ceux avec qui j'ai collaboré pour la réalisation des différentes parties de cette thèse, et plus particulièrement : Olivier Coriton et Virginie Huteau de l'INRA du Rheu pour leur aide précieuse en cytogénétique moléculaire ; Spencer Brown et Olivier Catrice de l'Institut des sciences du végétal de Gif-sur-Yvette pour leur aide en cytométrie en flux ; Rémy Pasquet de l'IRD (Nairobi, Kenya) pour nous avoir fourni des graines de *Lupinus princei* ; le professeur Bogdan Wolko et son équipe de l'Institut de génétique des plantes de Poznań (Pologne) pour avoir sélectionné et fourni un BAC extrait de la première banque génomique de lupin ; Higinio Pascual et Alberto Navarro-Perris du Centre de ressources végétales de Valence pour nous avoir fourni des graines de la nouvelle espèce récemment découverte en Espagne, *Lupinus mariae-josephi*.

Merci, bien sûr, à tous les membres de l'équipe EGS/MOB (enseignants, chercheurs, personnels, étudiants) ainsi qu'à tout le personnel de l'UMR Ecobio pour l'aide et le soutien multiformes qu'ils m'ont apportés tout au long de ce travail. J'adresse des remerciements particuliers aux stagiaires de licence ou master qui ont participé à certaines parties de ce travail : Christophe Biteau, Anis Bessadok, Émilie Robin et Mira Markova.

Table des matières

Introduction

« Partons donc : nous voulons, les deux, la même chose.
Toi, tu seras le chef et le guide et le maître. »
Et sur ce, reprenant la marche interrompue,
J'entrai dans le pénible et sauvage chemin.

Dante Alighieri, *La Divine comédie*, L'Enfer, chant II.

LA VARIATION de la taille des génomes est l'un des principaux processus impliqués dans la dynamique évolutive des génomes (X. Zhang & Wessler, 2004 ; Bennett & Leitch, 2005b ; Bennetzen *et al.*, 2005 ; Petrov & Wendel, 2006). Cette variation n'est pas corrélée au degré de complexité des organismes (*C-value paradox* de Thomas, 1971). Chez les organismes eucaryotes, la taille des génomes varie considérablement et peut atteindre une grande magnitude (Grover *et al.*, 2004). Chez les angiospermes la magnitude observée est d'au moins 2 000 entre les plus petits et les plus grands génomes (Bennett *et al.*, 2000 ; Greilhuber *et al.*, 2006). Des variations aussi spectaculaires soulèvent la question de la nature et de l'impact pour les organismes, en termes évolutifs et adaptatifs, de l'accumulation ou de la perte d'ADN.

Chez les plantes, l'adaptation, l'évolution et la spéciation s'accompagnent souvent d'une variation plus ou moins importante de la taille des génomes[1], l'accumulation ou la perte d'ADN répétitif étant la première cause de cette variation (Petrov & Wendel, 2006, pour revue). En dehors de certains processus, tels les duplications (polyploïdie, duplications segmentaires) et les recombinaisons génomiques, ce sont les éléments transposables (ET) qui constituent l'essentiel de l'ADN répété (R. B. Flavell *et al.*, 1977 ; Barakat *et al.*, 1997 ; SanMiguel & Bennetzen, 1998).

Ainsi, les éléments transposables, et plus particulièrement les rétrotransposons — c'est-à-dire les éléments de classe I ; voir Vicient, Jääskeläinen *et al.* (2001) ; Capy

1. *Annals of Botany*, Jackson, M. (éd.), numéros spéciaux *82* (1998) et *95* (2005).

(2005) ; Neumann *et al.* (2006) ; Piégu *et al.* (2006) ; Zuccolo *et al.* (2007) — constituent une composante majeure des génomes des plantes (Bennetzen, 2002 ; Bennetzen *et al.*, 2005 ; Vitte & Panaud, 2003). Lorsqu'elle n'est pas létale, leur dynamique évolutive (structurale et fonctionnelle) va générer de la diversité génétique, aux conséquences diverses pour l'adaptation, l'évolution et la diversification des organismes (McClintock, 1984 ; Brandt *et al.*, 2005 ; Biémont & Vieira, 2006). Les recherches récentes indiquent de plus en plus leur implication dans les réponses à divers stress environnementaux (Wessler *et al.*, 1995 ; Grandbastien, 1998 ; Quattrocchio *et al.*, 1999 ; Kalendar *et al.*, 2000 ; Jiang *et al.*, 2003 ; Pourtau *et al.*, 2003) ou génomiques (Waugh O'Neill *et al.*, 1998 ; B. Liu & Wendel, 2000). Une variation affectant les traits liés à la reproduction peut avoir des conséquences déterminantes sur l'isolement reproductif et conduire éventuellement à un processus de spéciation (Bennetzen, 2002).

Dans ce contexte, nous nous proposons d'examiner le rôle des éléments transposables dans la variation de la taille des génomes chez des légumineuses du genre *Lupinus* (fabacées), en relation avec leur diversication et leur adaptation à des conditions éco-géographiques variées.

Le genre *Lupinus*, regroupant 200 à 500 espèces, présente deux centres majeurs de diversité : l'*Ancien Monde* et le *Nouveau Monde*, ce dernier regroupant plus de 90 % des espèces. Les lupins du Nouveau Monde sont annuels ou pérennes (herbacés ou ligneux), autogames ou allogames, avec des nombres chromosomiques de $2n = 36$, $2n = 48$, ou $2n = 52$ (Dunn, 1984 ; Planchuelo, 1984 ; Maciel & Schifino-Wittmann, 2002 ; Conterato & Schifino-Wittmann, 2006 ; Camillo *et al.*, 2006). Les lupins de l'Ancien Monde sont tous annuels autogames, avec des nombres chromosomiques très variables allant de $2n = 32$, 36, 38, 40, 50 et 52 (Plitmann & Pazy, 1984 ; Carstairs *et al.*, 1992 ; Gladstones, 1998). De façon générale, l'ensemble des lupins est considéré comme une série paléopolyploïde (Pazy *et al.*, 1977 ; W. Williams *et al.*, 1980 ; Dunn, 1984 ; Wolko & Weeden, 1989). Les lupins sont exploités pour la richesse de leurs graines en protéines, leur capacité à enrichir naturellement en azote les sols pauvres, et leur exceptionnel pouvoir d'épuration des eaux et sables contaminés par les métaux lourds et les pesticides comme l'atrazine[2] (Bonvallot, 2004).

Les lupins euro-méditerranéens et africains constituent un groupe particulier en terme d'évolution de taille du génome, avec des quantités d'ADN pouvant varier de 1 à 2,5 pg[3], y compris entre espèces ayant le même nombre chromosomique (Naganowska *et al.*, 2003 ; Biteau, 2004). Par exemple, une variation notable de la taille du génome est observée entre les espèces méditerranéennes à $2n = 52$ chromosomes, *Lupinus luteus* (2,42 pg) et *L. micranthus* (1,0 pg) ; ou encore entre des espèces africaines à $2n = 38$ chromosomes, telles que *L. atlanticus* (1,7 pg), endémique du Sud marocain, et *L. princei* (1,0 pg) endémique du Kenya (Aïnouche & Bayer, 1999 ; Aïnouche *et al.*, 2004). Ce système représente ainsi un modèle naturel intéressant offrant la possibilité

2. L'atrazine est un herbicide appartenant à la famille chimique des triazines, principalement utilisé pour maîtriser les mauvaises herbes graminoïdes. Très utilisé en Bretagne jusqu'en 2001 et interdit depuis en France (Briand, 2006).
`http://www.ccme.ca/sourcetotap/atrazine.fr.html`

3. Un picogramme (pg) équivaut à 978 millions de paires de bases (Bennett *et al.*, 2000).

d'appréhender les mécanismes mis en jeu dans l'évolution des génomes de plantes, et d'explorer leurs implications dans l'adaptation et la spéciation.

Dans ce travail, un intérêt particulier sera accordé aux espèces adaptées à des conditions écologiques très contrastées, et présentant des différences remarquables de la taille de leurs génomes, en vue d'évaluer : 1) la diversité des rétrotransposons et leur implication dans les différences de taille des génomes ; 2) la dynamique évolutive de ces éléments au sein des génomes de lupins ; et 3) le degré de corrélation entre les différences de taille de génomes observées et la diversification des lupins (phylogénie) dans des conditions éco-géographiques contrastées.

Pour cela, notre démarche a consisté à :

- développer de nouveaux jeux de données moléculaires (espaceurs de l'ADN nucléaire ribosomiques, et gène *Symbiotic Receptor Kinase*) pour améliorer la phylogénie du genre *Lupinus*, en vue de préciser le cadre évolutif indispensable à la compréhension de l'histoire des lupins et à l'interprétation de la dynamique évolutive des éléments transposables ;
- évaluer la diversité et la dynamique évolutive des rétrotransposons de type Ty1/*copia* et Ty3/*gypsy* dans le genre *Lupinus* et quelques représentants de leurs alliées génistoïdées, en nous appuyant sur l'analyse phylogénétique de leurs séquences codant la transcriptase inverse ;
- évaluer l'importance relative de ces éléments dans des couples d'espèces remarquables pour leurs différences de taille de génome par des méthodes de cytogénétique moléculaire (hybridation *in situ*) et de PCR semi-quantitative ;
- estimer l'impact des éléments transposables à une échelle local du génome des lupins à partir du séquençage et de l'annotation d'un BAC de 116 kb contenant le gène d'intérêt *Symbiotic Receptor Kinase*, et de sa comparaison à des régions orthologues de différentes lignées de légumineuses modèles.

Dans une première partie, nous présenterons une revue bibliographique générale sur la dynamique des génomes (p. 11), le rôle des éléments transposables (p. 23), ainsi qu'une présentation du modèle *Lupinus* (p. 37). La deuxième partie sera consacrée à la description du matériel biologique (p. 57) et à la présentation des méthodes utilisées pour répondre aux objectifs posés (p. 71). Les résultats seront exposés dans une troisième partie, selon trois volets correspondants à nos trois axes d'étude :

1. phylogénie moléculaire du genre *Lupinus* (p. 95),
2. variation de la taille des génomes et éléments transposables (p. 135),
3. analyse génomique comparative de la région *SymRK* de *Lupinus angustifolius* et comparaison avec celles d'autres papilionacées (p. 153).

Enfin, une synthèse des résultats des différentes parties sera présentée à la fin du mémoire dans une discussion générale.

Première partie

Dynamique des génomes, rôle des éléments transposables et présentation du modèle biologique

1

Dynamique des génomes

« Il est hors de doute que l'étude systématique, de la teneur absolue du noyau en acide désoxyribonucléique, à travers de nombreuses espèces animales, puisse fournir des suggestions intéressantes en ce qui concerne le problème de l'évolution. »

Vendrely & Vendrely (1950)

DANS cette première partie, nous ferons un bilan bibliographique sur la dynamique des génomes eucaryotes et notamment les variations de taille rencontrées d'une espèce à l'autre, sur les causes possibles de ces variations et sur le rôle des éléments transposables dans ces variations de taille du génome.

Peser un génome La taille des génomes[1] eucaryotes est mesurée suivant un critère standardisé : la valeur C[2]. Cette valeur est égale à la masse du génome haploïde — voir Greilhuber *et al.* (2005) pour une définition des différentes valeurs C employées — et est exprimée en picogrammes (pg)[3]. La relation utilisée pour convertir la masse d'ADN en un équivalent en nombre de paires de bases, est basée sur les approximations suivantes :

1. L'expression « taille de génome » est souvent attribuée de façon erronée à Ralph Hinegardner (Greilhuber *et al.*, 2005). En effet, il utilise à deux reprise cette expression (Hinegardner, 1968, 1976), mais pas dans son sens moderne. La taille de génome désigne pour lui le nombre de gènes. Ce sont Wolf *et al.* (1969) qui devraient être crédités pour la première utilisation de l'expression « taille de génome » pour désigner la totalité de l'ADN contenu dans le noyau, définition utilisée aujourd'hui.

2. L'expression « C-value » a été créée par Hewson Swift (1950) — généticien expert en microscopie électronique, il participe à la fondation en 1960 de l'*American Society for Cell Biology* —, et se réfère à la « constance » de la quantité d'ADN dans un génotype donné. Il se base lui-même sur un article de deux biologistes français qui notent « une remarquable *constance* dans la teneur en acide désoxyribonucléique du noyau de toutes les cellules et chez tous les individus dans une même espèce animale » (Vendrely & Vendrely, 1948).

3. Un picogramme représente un millième de milliardième de gramme (10^{-12} g). Plus rarement, on trouvera des masses exprimées en Dalton (Da), un nucléotide équivalant à environ 330 Da. Enfin pour rappel, 1 kb, 1 Mb et 1 Gb représentent respectivement un millier, un million et un milliard de nucléotides.

- l'ADN est dissocié (les deux brins sont séparés) ;
- les masse atomiques ont pour valeurs :
  - $H = 1,00797$, $C = 12,0115$, $N = 14,0067$, $O = 15,9994$, $P = 30,9738$;
- 1 dalton $= 1,65979 \times 10^{-24}$ g ;
- le ratio AT : GC est de 1.

La relation entre masse (en pg) et nombre de paires de bases (en milliards) est la suivante (Bennett *et al.*, 2000) :

$$\text{Nombre de paires de bases} = \text{masse} \times 0,980$$

La mesure de la valeur C n'est pas triviale. La méthode la plus fréquemment utilisée, la cytométrie en flux, se base sur un agent intercalant fluorescent qui se fixe sur l'ADN et permet de colorer le génome. La quantité d'intercalant fixée, et donc la valeur C inférée, dépend du niveau d'empaquetage de la chromatine qui lui-même dépend du niveau de stress de l'organisme, de l'état énergétique de la cellule ou de la température de la pièce (Nardon *et al.*, 2003 ; Doležel & Bartoš, 2005). Il a été récemment confirmé que certains métabolites secondaires, comme les anthocyanines, influent fortement (35-70 %) sur la fixation de l'agent intercalant (Greilhuber, 1998 ; Bennett *et al.*, 2008). En raison de ces difficultés et de la variabilité naturelle, les valeurs C ne sont considérées comme fiables que si elles ont été obtenues indépendamment par plusieurs laboratoires (Greilhuber, 2008).

Les valeurs extrêmes rapportées à ce jour pour les eucaryotes sont 2,8 Mb pour *Encephalitozoon cuniculi* (Biderre *et al.*, 1995) et plus de 690 000 Mb pour la diatomée *Navicola pelliculosa*[4] (Cavalier-Smith, 1985, cité par Hawkins *et al.*, 2006). À noter également la taille du génome d'*Ostreococcus tauri*, le plus petit représentant de la Lignée verte connu, avec un diamètre cellulaire moyen de seulement 0,8 µm et un génome de 12,56 Mb réparties sur 20 chromosomes, un chloroplaste et plusieurs mitochondries (Derelle *et al.*, 2006).

Du paradoxe à l'énigme de la valeur C L'expression « paradoxe de la valeur C » a été utilisée par Charles Thomas (1971) pour décrire le fait que les grandes variations de valeur C constatées ne sont pas corrélées à la complexité des organismes ou au nombre de gènes contenus dans leurs génomes. Par exemple, bien que le génome de l'orge (4 900 Mb) soit 11 fois plus grand que celui du riz (440 Mb), les deux plantes possèdent un nombre comparable de gènes (Bennetzen & Kellogg, 1997). En effet, cette différence était incompréhensible dans le contexte de l'époque puisque la découverte de l'ADN non-codant et de son importance dans les génomes eucaryotes ne se fera que dans la première moitié des années 1970. En 2001, T. Ryan Gregory propose l'expression « énigme de la valeur C » et quatre questions scientifiques gravitant autour de celle-ci :

1. Quel type d'ADN non-codant rencontre-t-on dans les génomes eucaryotes, et en quelle proportion ?

4. Cette valeur concernant la diatomée *Navicola pelliculosa* est contestée. Voir Gregory, 2005a pour plus de détails.

2. D'où vient cet ADN non-codant, et comment passe-t-il d'un génome à l'autre ?
3. Quels effets, voire quelles fonctions, a cet ADN non-codant ?
4. Pourquoi certaines espèces présentent-elles des génomes compacts alors que d'autres ont d'énormes quantités d'ADN non-codant ?

Trois bases de données ont été créées avec comme objectif la centralisation des valeurs C disséminées dans la littérature (Gregory *et al.*, 2007) : *Plant DNA C-values Database*[5], *Animal Genome Size Database*[6] et *Fungal Genome Size Database*[7]. Le découpage de ces bases de données reprend la division du monde vivant en cinq Règnes proposée par Whittaker (1969). Bien que pratique, cette structuration néglige une partie des avancées de la biologie moléculaire et la refonte de la phylogénie des eucaryotes en six super-règnes qui en a découlé — Opisthocontes, Amoebozoaires, Lignée verte, Rhizariens, Chromoalvéolés et Excavobiontes (Keeling *et al.*, 2005 ; Parfrey *et al.*, 2006 ; Yoon *et al.*, 2008 ; Hampl *et al.*, 2009). Néanmoins, ces bases ont le mérite d'exister et leur examen à la lumière de la phylogénie du monde vivant permet de nuancer « le paradoxe de la valeur C ». En effet, bien que la variation de taille des génomes soit considérable entre organismes de même niveau de complexité, il se dégage clairement une tendance générale liant la taille et les caractéristiques des génomes à l'augmentation de la complexité des organismes (Lynch, 2007b).

1.1 Variation de la taille du génome chez les angiospermes

À l'heure actuelle, les tailles de génomes ont été mesurées pour environ 5 000 angiospermes, soit environ 2 % des 230 000 espèces connues. L'objectif — défini en 2003 au cours du *Second Plant Genome Size Workshop and Discussion Meeting* — est d'aboutir en 2009 à une couverture de 75 % des familles et 10 % des genres d'angiospermes.

D'après les données de la *Plant DNA C-values Database*, les angiospermes présentent une variabilité de taille de génome très importante (voir Fig. 1.1 page suivante), allant de 0,1 pg chez *Fragaria viridis*[8] à 127,4 pg chez *Fritillaria assyriaca* (I. J. Leitch *et al.*, 1998 ; Soltis *et al.*, 2003 ; I. J. Leitch *et al.*, 2005). La taille de génome moyenne est de 6,51 pg mais l'importance de l'écart type ($\sigma = 9,92$) indique clairement que la répartition des valeurs C ne suit pas une loi normale mais une loi de puissance (Clauset *et al.*, 2007) ou loi de Pareto[9]. Ce qui illustre et renforce le caractère énigmatique de la variation de la taille des génomes chez des organismes de même degré général de complexité structurale et fonctionnelle. Cette variation soulève à la fois la question des forces et des mécanismes à l'origine de l'évolution de l'architecture des génomes, et de sa signification en termes évolutifs et adaptatifs.

5. `http://data.kew.org/cvalues/homepage.html`
6. `http://www.genomesize.com`
7. `http://www.zbi.ee/fungal-genomesize/`
8. *Fragaria viridis* Duchesne : le fraisier vert ou fraisier des collines est une plante vivace de la famille des rosacées.
9. Vilfredo Pareto (1848-1923), sociologue et économiste italien. En 1906, il observe que vingt pour cent de la population possède quatre-vingt pour cent de la propriété en Italie, observation à l'origine de la loi qui porte son nom. Ce type de distribution est également appelé « distribution à longue traîne »

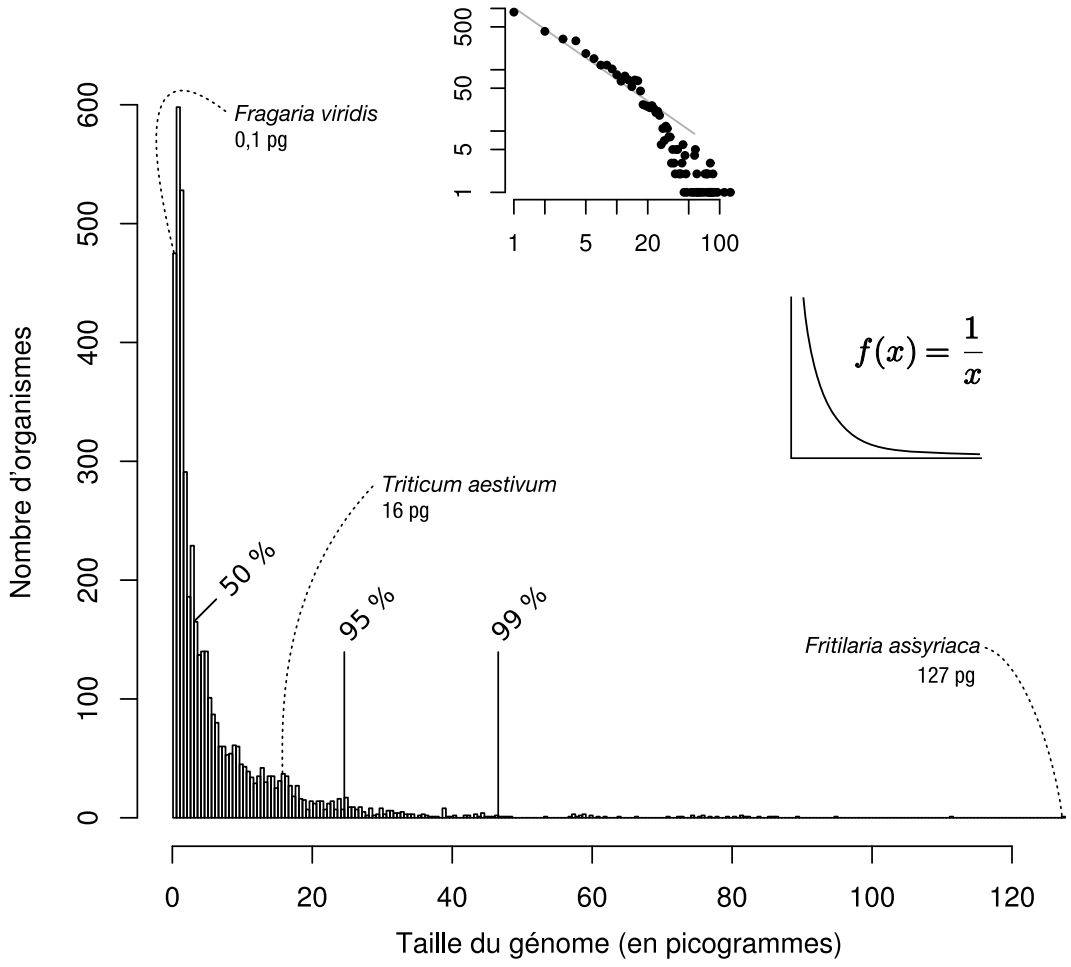

FIG. 1.1 — Répartition des valeurs 1C mesurées chez 4 427 angiospermes, données extraites de l'*Angiosperm DNA C-values database* (Bennett & Leitch, 2004). La répartition des valeurs suit une loi de Pareto ou « loi à longue traîne » de la forme $f(x) = x^{-1}$. La projection logarithmique montre cependant une dispersion importante de la fréquence des grandes tailles de génomes. Les valeurs sont comprises entre 0,1 pg (*Fragaria viridis*) et 127,4 pg (*Fritillaria assyriaca*), la largeur de classe est de 1 pg. Près de 50 % des plantes ont un génome de taille supérieure à 2,9 pg et 5 % des plantes ont une taille de génome supérieure à 23 pg.

Des données non recensées dans la base de données de Kew (Greilhuber *et al.*, 2006), font passer la variation de taille de génome chez les angiospermes de 800 × à près de 2 000 ×. En effet, les valeurs C enregistrées chez certains représentants de la famille des Lentibulariaceae (Asteridae) — plantes carnivores hautement spécialisées — sont extrêmement faibles. Trois taxons de cette famille possèdent des génomes d'une taille inférieure à celle que l'on rencontre chez le fraisier : *Genlisea margaretae* avec 63 Mb, *Genlisea aurea* avec 64 Mb et *Utricularia gibba* avec 88 Mb.

Ces variations se retrouvent à l'échelle des familles. Chez les brassicacées par exemple, les tailles de génomes sont en général faibles : 0,63 pg en moyenne, avec un génome ancestral estimé à 0,5 pg. Cependant on constate des variations de 16, 2× au sein de la famille puisque certaines tribus (Anchonieae et Physarieae) semblent avoir accumulé de l'ADN bien plus rapidement que les autres (Lysák *et al.*, 2009). Chez les fabacées également les tailles de génome sont en général très inférieures à la moyenne des angiospermes (6,5 pg), avec toutefois des variations de 35× entre *Lablab niger* (0,4 pg) et *Vicia faba* (13 pg) (Young *et al.*, 2003).

Au niveau générique, il existe des exemples de variations importantes. Le genre *Genlisea*, cité plus haut comme abritant le plus petit génome d'angiosperme connu (*Genlisea margaretae* avec 63 Mb), contient également l'espèce *Genlisea hispidula* (1 510 Mb), soit une variation de 24×. D'autres cas de variations importantes au sein du même genre ont été rapportés dans différentes lignées d'angiospermes, comme par exemple chez *Gossypium* (Wendel *et al.*, 2002), *Hordeum* (Jakob *et al.*, 2004) ou chez les orobanches (Weiss-Schneeweiss *et al.*, 2005). Chez *Vicia* (Fabaceae), qui comprend plus de 160 espèces, les tailles de génome sont comprises entre 1,9 et 14,4 pg, soit une variation de 8× (Neumann *et al.*, 2006). Des cas de variation de taille de génome au niveau infra-spécifique ont également été rapportés chez plusieurs espèces : 10 % chez *Arabidopsis thaliana* (Schmuths *et al.*, 2004), de 20 à 24 % chez *Dasypyrum villosum* (Cremonini *et al.*, 1994), 32 % chez *Helianthus annuus* (Michaelson *et al.*, 1991) et 40 % chez *Zea mays* (Rayburn *et al.*, 1985)[10]. Des données récentes indiquent des variations de taille de génome de 11,7 % entre graines d'une même inflorescence et de 18,8 % à l'échelle d'une même population de *Festuca pallens* (Šmarda *et al.*, 2008).

Une telle variation de la taille des génomes soulève plusieurs questions quant à la nature de l'ADN accumulé, aux forces et mécanismes à l'origine de cette variation, et ses conséquences sur les organismes en termes évolutifs et adaptatifs.

1.2 Influence et causes des variations de taille de génome

1.2.1 Relation entre taille de génome et traits d'histoire de vie

Chez les eucaryotes, la taille du génome — c'est-à-dire la quantité d'ADN par noyau — s'étend sur près de cinq ordres de grandeurs (10^5). Cette importante variation a très tôt été corrélée positivement avec le volume du noyau (Baetcke *et al.*,

10. Ces exemples de variations infra-spécifiques sont contestées par Greilhuber (1998, 2005) et Bennett *et al.* (2008).

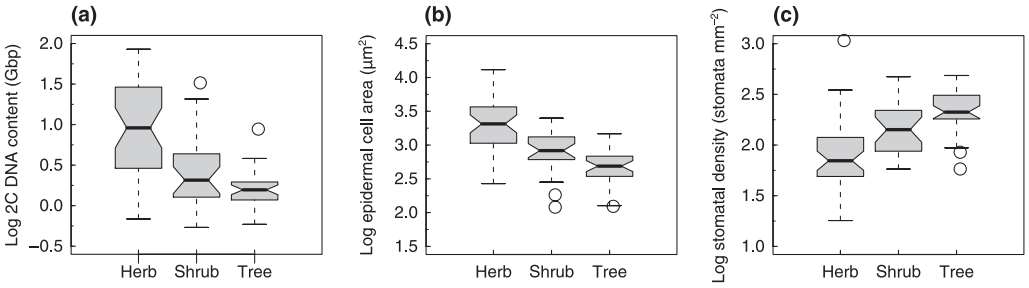

FIG. 1.2 — Relations entre le port de la plante et a) la taille de son génome diploïde, b) la surface de ses cellules épidermiques et c) la densité de ses stomates (valeurs mesurées pour 41 espèces herbacées, 26 espèces buissonnantes et 34 espèces arborescentes). Le trait noir représente la médiane. La boîte grisée représentent les premiers et troisièmes quartiles et les moustaches représentent les 5es et 95es centiles. Les cercles blancs représentent les individus aberrants. Les différences constatées entre les herbes, les buissons et les arbres sont significatives pour les traits a, b et c. Par rapport aux plantes herbacées, les arbres et les buissons ont des tailles de génome inférieures, des cellules plus petites (données non-représentées) et une plus grande densité en stomates. Figure modifiée d'après Beaulieu *et al.* (2008).

1967 ; Jovtchev *et al.*, 2006) ainsi que le volume cellulaire (Mirsky & Ris, 1951 ; Bennett, 1972 ; Price *et al.*, 1973), et corrélée négativement avec la durée du cycle cellulaire (Van't Hof & Sparrow, 1963 ; Evans *et al.*, 1972 cités par Cavalier-Smith, 1978 ; Cavalier-Smith, 2005 ; Francis *et al.*, 2008[11]). Une corrélation entre la taille du génome et les traits suivants à également été suggérée pour les macro-organismes : la taille du corps, le taux métabolique, la vitesse de croissance, la complexité, les stratégies de reproduction (Wright *et al.*, 2008), la distribution géographique, la taille relative du cerveau (chez les perroquets, Andrews & Gregory, 2009) et le risque d'extinction (pour revue, voir Kellogg & Bennetzen, 2004 ; Bennett & Leitch, 2005a ; Gregory, 2005b).

Chez les angiospermes, une corrélation positive a été suggérée entre la taille du génome, la taille des grains de pollen (Misset & Gouret, 1996), et la masse des graines (Bennett, 1972 ; K. Thompson, 1990 ; Knight & Ackerly, 2002 ; Gaut & Ross-Ibarra, 2008). Une étude de Beaulieu *et al.* (2007) est cependant venue nuancer cette dernière hypothèse et montrer que la relation entre taille du génome et taille des graines n'est pas directe et présente des effets de seuils. L'année suivante, les mêmes auteurs (Beaulieu *et al.*, 2008 ; Knight & Beaulieu, 2008) ont par contre confirmé la corrélation positive entre le type de port de la plante — herbacé, buissonnant ou arborescent — et la taille du génome (voir Fig. 1.2). Les arbres ont donc tendance à avoir des génomes plus petits et une variance plus faible que les herbacées ou les buissons. Ces résultats appuient l'hypothèse d'une relation entre la taille de génome et la taille *minimale* de la cellule (Bennett, 1972 ; Gregory, 2001). En renversant cette relation il devient possible, à partir des empreintes fossiles de tissus végétaux, d'évaluer d'anciennes tailles de génome et de retracer l'évolution de ce trait (I. J. Leitch *et al.*, 2007).

11. Les effets dus à la taille de génome semblent devenir très marqués au-delà de 25 pg (Francis *et al.*, 2008). Ce qui, d'après les statistiques sur la banque de donnée de Kew, concerne moins de 5 % des angiospermes.

Des stomates de petite taille et plus nombreux (c'est-à-dire plus denses) permettent à la plante de mieux contrôler les échanges gazeux et les pertes hydriques (Hetherington & Woodward, 2003). Beaulieu *et al.* (2008) montrent qu'un gros génome n'est pas compatible avec une forte densité en stomates, les plantes à gros génomes étant peu fréquentes dans les environnements chauds et secs (Knight & Ackerly, 2002). De plus, une grande densité en stomates semble nécessaire pour déplacer l'eau et les nutriments sur une hauteur importante (Woodward, 1998). Une augmentation de la taille de génome — entraînant une baisse de la densité en stomates — aurait donc un impact négatif sur les arbres. Cette hypothèse est compatible avec le fait que la polyploïdie (définition p. 19) soit rare chez les angiospermes arborescentes (Meyers & Levin, 2006)[12], il pourrait donc s'agir d'une contrainte écologique limitant l'augmentation de taille des génomes (Knight *et al.*, 2005).

A. R. Leitch & Leitch (2008) font remarquer que les plantes à grands génomes ont des besoins en nutriments supérieurs, particulièrement en nitrates et phosphates nécessaires à la fabrication des acides nucléiques. Ces plantes sont donc contre-sélectionnées lorsqu'elles poussent sur des sols pauvres, ce qui favorise à terme les plantes capables d'éliminer l'ADN en excès[13]. Enfin, concernant la relation entre taille du génome et stratégie reproductive, les plantes dites « envahissantes », à reproduction importante et rapide, semblent avoir des génomes plus petits que ceux des autres plantes (Bennett *et al.*, 1998). Cette tendance a été confirmée chez certaines plantes issues de la sélection artificielle, telles que le maïs (Laurie & Bennett, 1985 ; Rayburn *et al.*, 1985 ; Rayburn & Auger, 1990, cités par Greilhuber, 2005). Chez *Pisum sativum* au contraire, aucune corrélation de ce type n'a pu être mise en évidence malgré un échantillonnage à l'échelle mondiale (Baranyi & Greilhuber, 1995, 1996 ; cités par Greilhuber, 2005).

Les caractères évoqués, taille des cellules ou nombre de stomates, sont des caractères micro-morphologiques. La liaison entre taille de génome et macro-caractères — masse des graines ou masse des feuilles par unité de surface — est plus difficile à établir (Jakob *et al.*, 2004 ; Knight & Beaulieu, 2008). Il semble toutefois que les plantes à grands génomes aient tendance à avoir de plus grosses graines (Gaut & Ross-Ibarra, 2008) mais aussi une photosynthèse plus faible et une croissance plus lente (Knight *et al.*, 2005 ; Knight & Beaulieu, 2008). L'influence de la taille de génome sur la vie de l'organisme est donc réelle, mais sa portée est difficile à évaluer et varie de façon très importante d'une lignée à l'autre (Lynch, 2007b).

Quels sont les mécanismes à l'origine des variations de taille de génome ?

12. La polyploïdie est également rare chez les autres spermatophytes (ginkgophytes, pinophytes, cycadophytes, gnétophytes).

13. Cette prédiction du contrôle que les besoins énergétiques imposent sur la taille du génome trouve une confirmation chez les vertébrés. Les trois groupes de vertébrés ayant évolué vers le vol battu — ptérosaures, oiseaux et chauves-souris — ont tous des tailles de génomes inférieures à la moyenne des lignées auxquelles ils appartiennent : archosauriens, pour les ptérosaures et les oiseaux (Organ *et al.*, 2007 ; Organ & Shedlock, 2009) et mammifères pour les chauves-souris (J. D. Smith & Gregory, 2009).

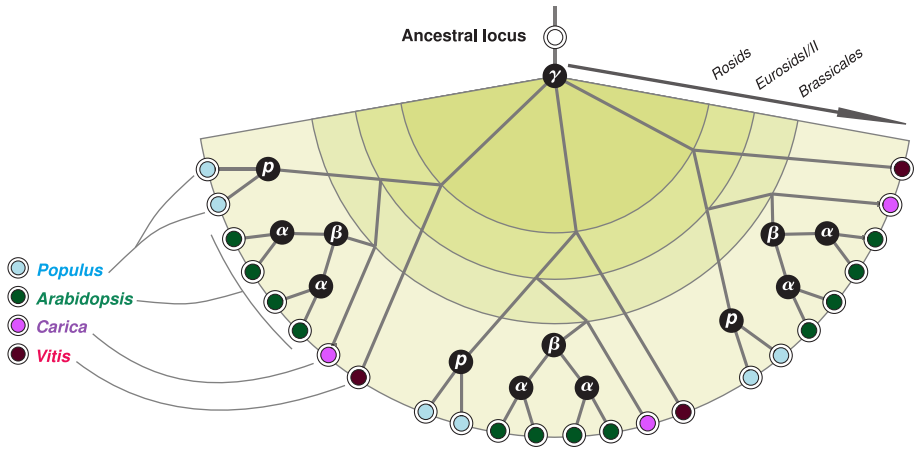

FIG. 1.3 — Vue idéalisée des gènes paralogues et orthologues chez *Populus trichocarpa*, *Arabidopsis thaliana*, *Carica papaya* et *Vitis vinifera*. Aucun gène n'est perdu suite à une polyploïdisation et les taux d'évolution sont supposés égaux pour toutes les branches. Les évènements de polyploïdisation sont représentés par des cercles noirs étiquetés α et β dans la lignée menant à *Arabidopsis* (The Arabidopsis Genome Initiative, 2000 ; Bowers *et al.*, 2003), *p* dans la lignée menant à *Populus* (Tuskan *et al.*, 2006) et γ chez l'ancêtre commun à ces quatre espèces (Jaillon *et al.*, 2007 ; Ming *et al.*, 2008). Modifié d'après Tang *et al.* (2008).

1.2.2 Causes et nature des variations de taille de génome

Les études de cinétique de réassociation de l'ADN dénaturé ont révélé que le génome des eucaryotes peut être grossièrement séparé en quatre fractions (Britten & Kohne, 1968) :

1. les séquences palindromiques capables de former des structures en « épingle à cheveux » ;
2. les séquences hautement répétées (généralement inférieures à 500 pb) ;
3. les séquences moyennement répétées (de quelques centaines à quelques milliers de paires de bases) ;
4. les séquences uniques.

La proportion génomique représentée par les séquences répétées est faible pour les taxons ayant un petit génome. Elle devient par contre très importante pour les taxons ayant un grand génome. C'est le cas pour *Citrus clementina* dont le génome de 367 Mb contient 12,5 % d'ADN répété (Terol *et al.*, 2008), et de *Secale cereale* dont le génome de 8 280 Mb contient 84 % d'ADN répété (Bartoš *et al.*, 2008).

L'ADN répété est donc à l'origine de l'essentiel des variations de taille de génome. Mais quels sont les mécanismes pouvant conduire à l'accumulation d'ADN répété ? L'augmentation de la valeur C peut se faire à deux échelles différentes : l'échelle locale (duplication d'une séquence de quelques paires de bases à quelques dizaines de milliers de paires de bases) et l'échelle globale (duplication du génome entier ou d'un chromosome) (Bennetzen, 2002).

Duplications génomiques et chromosomiques La duplication génomique, ou polyploïdie, consiste au rassemblement dans un même noyau d'au moins deux génomes. Les génomes peuvent être identiques (autopolyploïdie) ou d'origines différentes (allopolyploïdie) (Otto, 2007, pour revue). Le séquençage de génomes complets et la génomique comparative ont permis de mettre en évidence des traces d'anciens évènements de polyploïdisation dans la plupart des lignées eucaryotes (Jaillon *et al.*, 2009). C'est le cas par exemple chez les vertébrés où au moins deux évènements ont eu lieu (Finn & Kristoffersen, 2007)[14]. Chez les angiospermes, un évènement au moins a eu lieu chez l'ancêtre commun des rosidées (Cui *et al.*, 2006 ; Jaillon *et al.*, 2007 ; Ming *et al.*, 2008). Les lignées descendantes ont ensuite suivi leur propre histoire, les brassicacées ayant par exemple connu deux autres évènements de polyploïdisation (voir Fig. 1.3 page ci-contre) et plus récemment pour certaines *Brassica*, un triplement du génome (Lysák *et al.*, 2005 ; Parkin *et al.*, 2005). La plupart des espèces considérées aujourd'hui comme des diploïdes seraient donc des paléopolyploïdes (Bowers *et al.*, 2003 ; De Bodt *et al.*, 2005 ; Fawcett *et al.*, 2009).

La polyploïdie est un phénomène courant chez les plantes, et il est possible d'observer un grand nombre de polyploïdes récents, particulièrement parmi les espèces cultivées (Wendel, 2000 ; Adams & Wendel, 2005 ; A. R. Leitch & Leitch, 2008). C'est par exemple le cas du soja *Glycine max*, une espèce autotétraploïde ($2n = 4x = 20$) ou celui du blé tendre *Triticum aestivum*, une espèce allohexaploïde ($2n = 6x = 42$). La formation d'individus polyploïdes est un évènement courant dans la nature, et ce processus représente un mécanisme de spéciation majeur chez les angiospermes (Peer *et al.*, 2009 ; Fawcett *et al.*, 2009 ; Wood *et al.*, 2009). Ce phénomène est-il contrôlé par la sélection naturelle ? En 2006, une expérience a été menée sur la stabilité de la condition non-diploïde chez la levure *Saccharomyces cerevisiae*, en créant artificiellement des populations haploïdes et tétraploïdes. Après 1 800 générations, une convergence vers le stade diploïde a été observée dans toutes les lignées, que l'expérience ait été menée en situation de stress ou pas. Il y a donc une pression de sélection s'exerçant sur le critère « taille du génome » favorisant un retour vers le stade diploïde (Gerstein *et al.*, 2006). Cependant, les tailles de populations et les temps de générations habituellement rencontrés chez les angiospermes sont très différents de ceux d'une culture de levure. Cette différence d'échelle, ainsi que de possibles avantages liés à la polyploïdie, pourraient expliquer le maintien d'espèces polyploïdes sur plusieurs millions d'années (Ma *et al.*, 2004 ; Meyers & Levin, 2006 ; Lim *et al.*, 2007).

La duplication chromosomique, ou hyperploïdie, se définit par la présence d'un ou plusieurs chromosomes surnuméraires. Chez les plantes, ces formes de trisomie existent mais sont rarement transmises à la descendance car elles induisent le plus souvent des phénotypes pathologiques.

14. En 1970, Susumu Ohno émet l'hypothèse qu'une série de duplications complètes de génomes a eu lieu tôt dans l'histoire des vertébrés et aurait permis l'augmentation de complexité structurale et comportementale qui a suivi (hypothèse « 2R »). Devenue très populaire, cette hypothèse est cependant contestée par Abbasi (2008), mais appuyée par des travaux plus récents de Santini *et al.* (2009).

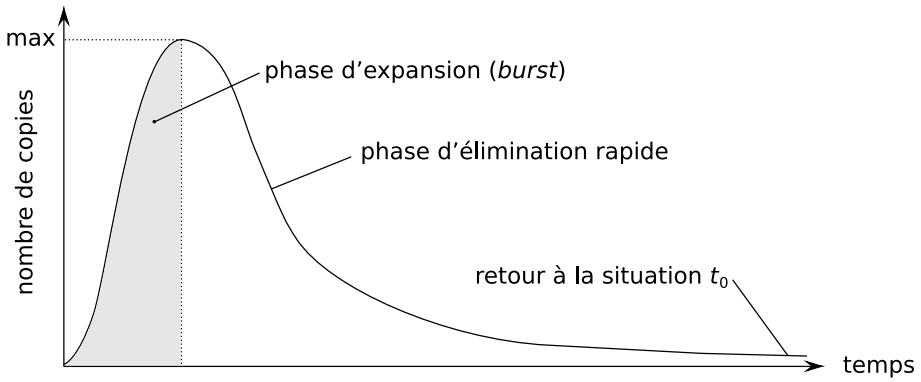

FIG. 1.4 — Représentation schématique d'un évènement de réplication d'éléments transposables. L'importance de la recombinaison non-homologue est corrélée avec le nombre de copies. La décroissance est donc initialement rapide, puis de plus en plus lente, et tend finalement vers un niveau proche de la taille de génome de départ.

Duplications locales Plusieurs catégories de répétitions existent sous le terme générique de duplication locale. Suite à l'étude du génome d'*Homo sapiens*, les duplications segmentaires ont été définies comme des régions répétées dont les copies partagent entre 90 et 100 % d'identité pour des tailles comprises entre 1 et 200 kb (Lander *et al.*, 2001). Mais cette définition fluctue en fonction de la taille du génome de l'organisme considéré. Environ 63 % du génome d'*Arabidopsis thaliana* — espèce considérée comme diploïde — correspondraient à des duplications segmentaires (Choisne *et al.*, 2004), mais il pourrait également s'agir des traces des trois évènements de polyploïdisation qu'a connu cette lignée (Bowers *et al.*, 2003). Une autre classe de répétitions de faible complexité, composées de motifs répétés, est présente dans les génomes eucaryotes : les satellites. Les satellites sont des répétitions de plusieurs dizaines de nucléotides, localisées le plus souvent dans les régions hétérochromatiques. Le satellite α des primates mesure 171 pb et représente, chez *Homo sapiens*, de 3 à 5 % de chaque chromosome (Lander *et al.*, 2001). Les mini-satellites sont des suites de monomères courtes (de 6 à 100 nucléotides) mais pouvant compter jusqu'à un million de copies et représenter jusqu'à 20 % de la masse d'un génome (Neumann *et al.*, 2006). Les micro-satellites quant à eux sont des suites de monomères, d'en général moins de 6 nucléotides, répétées en tandem plusieurs dizaines de fois.

Transposition La transposition s'avère être la cause principale des variations de taille de génome en dehors de la polyploïdie. En effet, les éléments transposables[15] sont, pour certains, capables de s'auto-répliquer et peuvent donc rapidement générer un grand nombre de copies d'eux-mêmes. Ils sont, avec la polyploïdie, un des facteurs inflationniste les plus efficaces et peuvent conduire eux aussi à un doublement de la taille de génome, comme constaté chez *Oryza australiensis* (Piégu *et al.*, 2006).

Les génomes sont-ils condamnés à l'obésité ? L'accumulation de répétitions est-elle

15. Les éléments transposables seront détaillés au chapitre 2 page 23.

irrémédiable ? C'est la question posée par Bennetzen & Kellogg (1997). Ne connaissant pas de mécanismes efficaces pour réduire la valeur C, ils postulaient néanmoins l'existence de ceux-ci et estimaient que les mécanismes inflationnistes et déflationnistes doivent vraisemblablement s'équilibrer. Quelques années plus tard, l'importance de la recombinaison non-homologue comme force déflationniste est démontrée par Shaked *et al.* (2001), Bennetzen (2002) et Devos *et al.* (2002). Sur un modèle animal, Petrov *et al.* (2000) démontrent que le taux de délétion de séquences répétées est 40 fois plus faible chez les criquets à gros génome que chez les drosophiles au génome plus petit, indiquant qu'il existe des différences d'intensité des mécanismes de délétion entre espèces qui expliqueraient l'existence de lignées à petit génome. Toujours selon Petrov (2001), le ou les mécanismes éliminant l'ADN répété n'ont pas nécessairement à être rapides. Un phénomène lent mais continu peut contrer des évènements majeurs mais rares. Cependant, d'après Ma *et al.* (2004), la recombinaison homologue ne peut pas ramener la taille du génome à exactement ce qu'elle était avant la phase d'expansion (voir Fig. 1.4 page ci-contre). Les mécanismes qui permettraient aux petits génomes de ne pas grossir ou d'éliminer l'ADN répété font donc toujours l'objet de débats (Lysák *et al.*, 2009 ; Hawkins *et al.*, 2009).

Les éléments transposables joue donc un rôle prépondérant dans les variations de taille de génomes (SanMiguel & Bennetzen, 1998 ; Wendel, 2000). Le chapitre suivant sera consacré à ces éléments, à leur diversité, à leur dynamique et à leur influence sur l'évolution des organismes.

2

Rôle des éléments transposables

« *Genome evolution is a continuous and dynamic process in which genomes expand and contract, duplicate in whole or in part and gain and lose genes at the same time as transposons amplify and are excised.* »

Kellogg & Bennetzen (2004)

LES GÉNOMES sont des structures dynamiques dont la taille et la composition varie (Bonnivard & Higuet, 2009). Chez l'homme, et c'est un des grands enseignements de l'ère de la génomique, la part du génome codant de manière directe pour des protéines est inférieure à 2 % du total. Comment passer du « tout génétique » de Jacques Monod (1970) à ce « moins de 2 % de codant » d'aujourd'hui ? Quelle est la nature et le rôle de cette masse d'ADN n'intervenant pas directement dans la production de protéines ? Les premiers grands projets de séquençage ont répondu à cette question : les éléments transposables (ET) constituent l'essentiel des génomes de grande taille.

Les éléments transposables — autoréplicatifs et mobiles — ont été découverts par Barbara McClintock (1950)[1] et il a fallu plusieurs dizaines d'années pour que leur importance soit réalisée par la communauté scientifique internationale (McClintock, 1983, discours de réception du prix Nobel). Ces éléments codent les instructions nécessaires à leur propre excision-réinsertion et sont présents chez la quasi-totalité des organismes étudiés à ce jour. Ils y forment la majeure partie de ce qui a été appelé l'« ADN poubelle » (*junk DNA*, Ohno, 1972, 1973)[2], et peut parfois représenter une importante partie du génome (voir Fig. 2.1 page suivante). Le statut de ces éléments

1. « Selon Barbara McClintock, l'expression différentielle des gènes résultait de leurs déplacements à travers le génome. Au cours de ces déplacements, ils étaient placés sous le contrôle de différents éléments régulateurs qui en modulaient l'expression. De fait, B. McClintock avait bien anticipé la découverte des gènes régulateurs. Cependant, la complexité du système sur lequel elle travaillait, certaines difficultés à communiquer ses résultats, et surtout l'importance majeure qu'elle attribuait au déplacement de gènes dans la régulation limitèrent l'impact de ses découvertes. » — Michel Morange (1994, p. 206).

2. Bien que péjorative, l'expression *junk DNA* utilisée par Susumu Ohno dans le titre de ses communications ne semble pas signifier un rejet de sa part. À aucun moment, il ne propose de se focaliser uniquement sur les régions codantes, et cherche au contraire à expliquer l'existence de cette masse d'ADN.

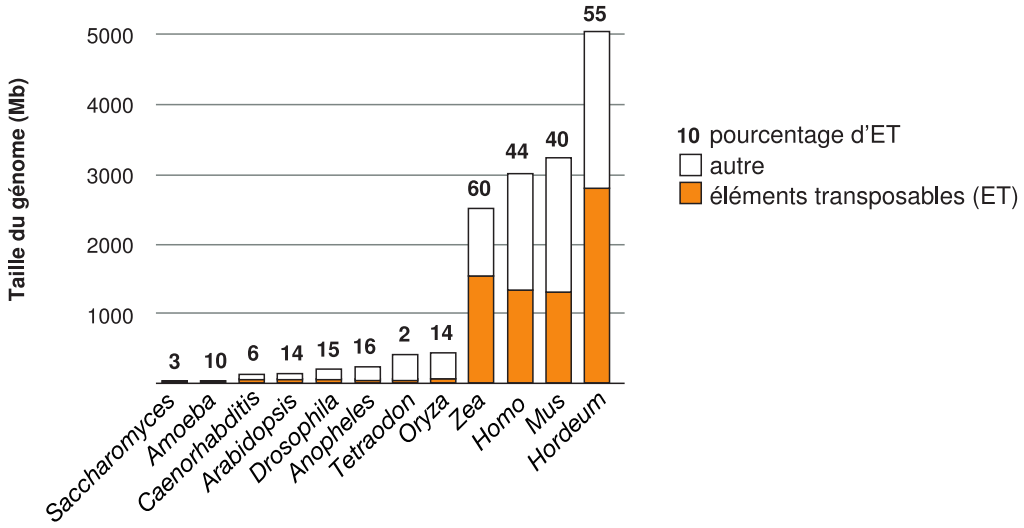

FIG. 2.1 — Taille du génome (en millions de paires de bases) et proportion d'éléments transposables chez douze eucaryotes d'après Kidwell (2002). Les chiffres au-dessus des barres indiquent le pourcentage d'éléments transposables. L'augmentation de taille de génome s'accompagne d'une augmentation de la teneur en éléments transposables (figure modifiée d'après Gaut & Ross-Ibarra, 2008).

a évolué avec l'accumulation de données : initialement considérés comme de l'ADN parasite ou égoïste[3] (Doolittle & Sapienza, 1980 ; Orgel & Crick, 1980, cités par Lönnig & Saedler, 1997), leur rôle moteur dans l'évolution des génomes et des organismes est aujourd'hui largement reconnu (McClintock, 1984).

2.1 Classification et nomenclature

La multiplication des projets de séquençage de génomes a permis la découverte d'un nombre grandissant d'éléments transposables, engendrant le besoin d'un système de classification. Les grandes familles décrites actuellement le sont sur la base de leurs mécanismes de transposition et sur la présence-absence de certaines fonctions ou séquences-cibles. Les éléments transposables se répartissent en deux grandes classes :

Classe I — Les éléments de la classe I, aussi appelés rétroéléments, se copient et s'insèrent *via* un intermédiaire ARN grâce à une machinerie enzymatique codée par l'élément lui-même (voir Fig. 2.2 page ci-contre). Par analogie, ce mécanisme est souvent résumé par l'expression « copier-coller ». Généralement de grande taille (quelques centaines à quelques milliers de paires de bases), ils peuvent représenter une part importante des génomes (voir par exemple SanMiguel & Bennetzen, 1998 ; Petrov & Wendel, 2006).

3. L'idée qu'un élément de génome puisse se comporter comme un parasite a été émise par Östergren (1945), et s'appliquait alors aux chromosomes B observés chez les plantes.

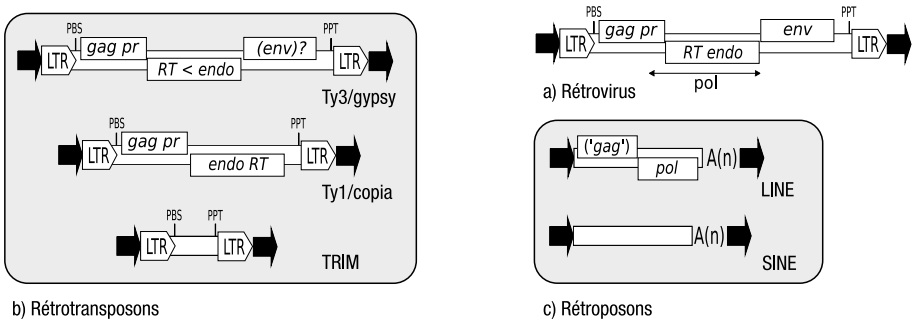

FIG. 2.2 — Les éléments transposables de classe I ou rétroéléments (texte et figure repris de Mhiri & Grandbastien, 2004).
(a) *Rétrovirus*. les rétrovirus s'insèrent dans le génome des hôtes infectés et se répliquent par l'intermédiaire d'un ARN. La forme intégrée du rétrovirus est bordée de deux longues répétitions terminales (LTR), dans lesquelles la transcription démarre et se termine, et entre lesquelles se trouvent les ORF codant pour les protéines nécessaires à la réplication virale. L'ARN matrice est encapsidé dans une particule virale dont le cœur est formé de protéines GAG. Associée à la membrane cellulaire de l'hôte, la protéine ENV permet ensuite au virion d'infecter une nouvelle cellule. L'ARN matrice est alors rétrotranscrit en une copie fille ADN par la fonction RT (transcriptase inverse), et la copie fille pénètre le noyau et s'insère dans le génome à l'aide de la fonction ENDO (endonucléase). Les sites PBS (*Primer binding Site*) et PPT (*Polypurine Tract*) sont impliqués dans l'amorçage de la transcription inverse.
(b) *Rétrotransposons*. Les rétrotransposons sont structurellement et fonctionnellement très proches des rétrovirus, à la différence qu'ils ne sont pas infectieux et que leur cycle d'amplification, qui passe par une VLP cytoplasmique (*Virus-Like Particle*), est purement intracellulaire. Toutefois, la frontière entre rétrovirus et rétrotransposons est floue. Les rétrotransposons se différencient en sous-classes, selon l'organisation des fonctions RT/ENDO. Les éléments TRIM sont de petits éléments dépourvus d'ORF, mais portant plusieurs caractéristiques des rétrotransposons (LTR, PBS, PPT).
(c) *Rétroposons*. Les rétroposons LINE, codants, et SINE, petits éléments non-codants qui utilisent vraisemblablement les fonctions portées par les LINE, sont dépourvus de LTR, mais s'amplifient également via un intermédiaire ARN.

Classe II — Les éléments de la classe II, appelés transposons à ADN, codent une enzyme « transposase » qui reconnaît les extrémités de l'élément, l'excise et l'insère ailleurs dans le génome (voir Fig. 2.3 page suivante). On parle dans ce cas de « couper-coller ». Plus courts que les éléments de classe I, leur mode de transposition limite leur importance en volume de séquence. Cependant, les éléments de classe II jouent un rôle évolutif important (Kazazian, 2004).

Les deux classes comportent des éléments non-autonomes : les SINE (*Short Interspersed Nuclear Element*) pour la classe I et les MITE (*Miniature Inverted-repeat Transposable Element*) pour la classe II. D'une taille comprise entre 100 et 1 000 nucléotides, ils ne contiennent pas de séquences codantes et utilisent donc les enzymes codées par des transposons autonomes pour se répliquer ou transposer.

Mhiri & Grandbastien (2004) décrivent les éléments de classe I comme une hiérarchie. Les rétrovirus sont les éléments les plus complexes, capables de s'encapsuler et de quitter le volume cellulaire. À l'opposé, les LINE (*Long Interspersed Nuclear Ele-*

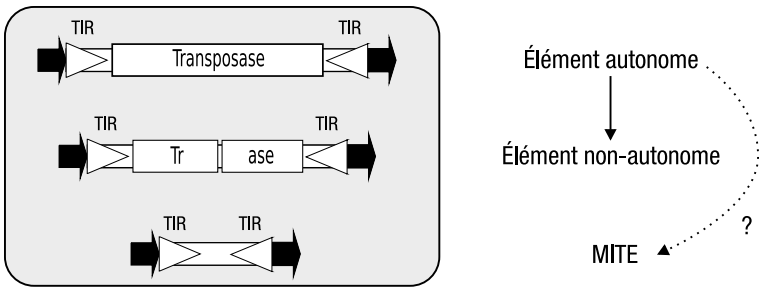

FIG. 2.3 — Les éléments transposables de classe II ou transposons à ADN (Mhiri & Grandbastien, 2004). Les transposons à ADN sont bordés de TIR (*Terminal Inverted Repeat*) caractéristiques de chaque famille, et se décomposent en éléments complets autonomes, porteurs d'une transposase active (et de plusieurs autres fonctions chez certains éléments complexes), et en éléments non-autonomes, transactivables par un élément autonome de la même famille. Les MITE sont dépourvus de séquences codantes, et sont probablement une forme particulière de dérivés non autonomes d'éléments complets (Casacuberta & Santiago, 2003). Tous les éléments transposables sont bordés de courtes duplications du gène cible (flèches noires) générées lors de l'insertion.

ment) sont des entités très simples, autonomes mais limitées au volume cellulaire. La polarisation de cette hiérarchie est débattue. Les rétrotransposons et les rétroposons sont-ils des dérivés de rétrovirus ayant perdu des modules fonctionnels, ou les rétrovirus sont-ils nés de formes plus simples par acquisition de fonctions ? C'est cette dernière piste qui est suivie par Wawrzynski *et al.* (2008). Selon eux, les LINE sont la classe d'éléments transposables la plus ancienne chez les eucaryotes. Ils seraient apparus suite à une acquisition de LTR (*Long Terminal Repeats*), les autres rétroéléments dériveraient des LINE par acquisitions successives de fonctions.

L'augmentation de l'effort de séquençage entraîne un besoin croissant d'annotation, supérieur à la disponibilité des experts. L'annotation doit donc être faite *in silico*, à l'aide de banques de séquences de référence et surtout à l'aide d'un système de classification. Le débat qu'entraîne la problématique de la classification des éléments transposables (Capy, 2005 ; Wicker *et al.*, 2007 ; Kapitonov & Jurka, 2008 ; Seberg & Petersen, 2009 ; Wicker *et al.*, 2009) n'est pas sans rappeler celui de la classification des organismes. Comme pour la classification du monde vivant, les partisans d'une systématique artificielle, mais fonctionnelle, affrontent les défenseurs d'une classification basée sur la phylogénie.

Wicker *et al.* (2007)[4] ont proposé une approche pragmatique pour faciliter le travail des annotateurs. Mais cette classification est critiquée par Seberg & Petersen (2009) et qualifiée d'artificielle. Le débat rejoint donc celui des taxonomistes, à la différence près que la méthodologie phylogénétique utilisée pour les organismes pourrait ne pas s'appliquer aux éléments transposables. Ceux-ci présentent des particularités qui rendent très difficile l'utilisation des schémas classiques : ce sont des séquences courtes, dont les différents modules peuvent évoluer à des vitesses très différentes, s'échanger, se perdre et s'acquérir à partir de sources phylogénétiquement lointaines. De plus les

4. `http://bioinformatics.org/wikiposon/doku.php`

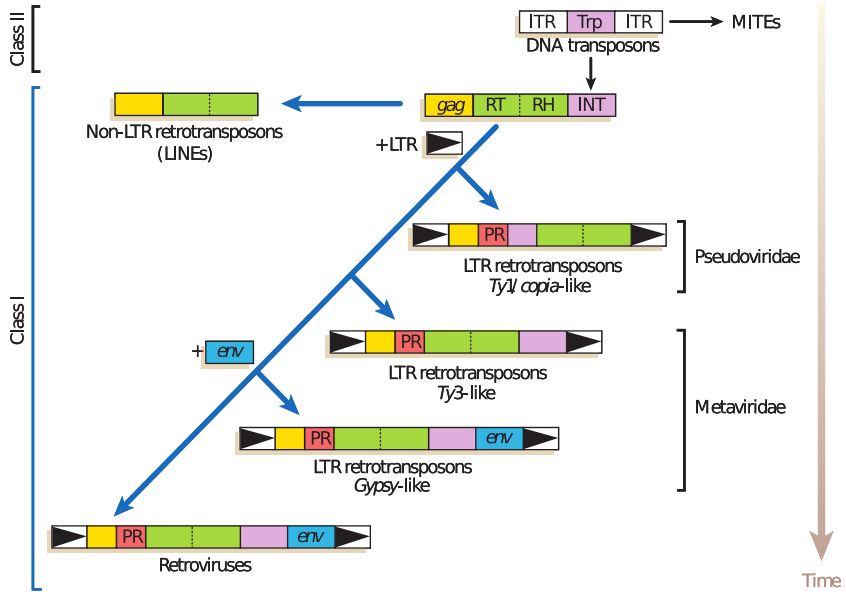

FIG. 2.4 — Proposition de classification des éléments transposables basée sur l'acquisition ou la perte de modules (Biémont & Vieira, 2006). Dans ce système, la fonction intégrase des éléments de classe I dérive d'un élément de classe II. Les différentes familles de rétroéléments émergent ensuite par acquisitions successives de fonctions. À noter que les groupes Pseudoviridae et Metaviridae sont paraphylétiques tels qu'ils sont représentés.

éléments transposables présentent des fréquences de recombinaisons et de mutations 10^3 fois plus rapide que les taux moyens constatés pour les gènes eucaryotes (Gabriel *et al.*, 1996).

Comment se traduit la modularité des éléments transposables ? Chez les rétrovirus, par exemple, la présence d'un troisième cadre ouvert de lecture (*Open reading frame* ou ORF) pour le gène enveloppe (*env*) n'est pas toujours une synapomorphie, c'est-à-dire un caractère partagé par un groupe et dérivant d'un ancêtre commun. Cet ORF a pu être acquis plusieurs fois, de façons indépendantes (Capy, 2005) et c'est probablement le cas du gène *env* des Errantivirus et de celui des éléments SIRE-1 (Hemivirus).

Biémont & Vieira (2006) proposent donc une autre classification basée sur l'acquisition de modules (voir Fig. 2.4). Dans leur proposition, l'évolution est polarisée — *via* une flèche temporelle — vers une augmentation de la complexité, ce qui se rapproche de la thèse de Wawrzynski *et al.* (2008). Biémont & Vieira ne donnent de précision que pour deux des nœuds de leur arbre de classification : le gain de LTR pour passer d'un rétroélément primitif à un rétrotransposon de type Ty1 et le gain d'une fonction *env* pour passer du type Ty1 au type *gypsy*. Les autres nœuds ne sont pas documentés.

2.2 Augmentations de taille de génomes liées aux éléments transposables

Les différentes familles de transposons ne sont pas toutes représentées au sein d'un même génome. On constate le plus souvent la domination d'une famille ou d'une sous-famille. Par exemple, le génome humain est dominé par la famille LINE 1 (17 % du génome), les éléments de type SINE (13 % du génome) et les éléments de la famille *Alu* (11 % du génome)[5]. À l'inverse, les génomes des grandes chauves-souris, qui sont en moyenne plus petits que ceux des autres mammifères (J. D. Smith & Gregory, 2009), semblent ne pas contenir d'éléments LINE-1. L'abondance des éléments SINE et des pseudogènes rétroprocessés[6] étant dépendante de l'élément LINE-1, son absence pourrait expliquer la relative petite taille des génomes des grandes chauves-souris (Cantrell *et al.*, 2008).

Chez les angiospermes, les grands génomes sont dominés par les rétrotransposons. SanMiguel & Bennetzen (1998) ont comparé la composition génomique de *Zea mays* ($n = 10$ et $4,9 < 2\text{C} < 5,5$ pg) et de *Sorghum bicolor* ($n = 10$ et $1,6 < 2\text{C} < 1,8$ pg), deux taxa cultivés s'étant différenciés il y a environ 16 millions d'années. En analysant une portion représentative du génome total, la région du gène *adh*1, les auteurs ont pu mettre en évidence la présence d'un plus grand nombre de rétrotransposons chez le maïs. Les rétrotransposons actuellement présents au sein du génome de *Zea mays* — et absents de celui de *Sorghum bicolor* — se sont donc multipliés et propagés depuis la séparation de ces deux espèces et représentent aujourd'hui près de 60 % du génome du maïs, soit un quasi triplement de la taille du génome de départ.

Le génome de l'espèce sauvage *Oryza australiensis* ($2C = 965$ Mb) est composé pour moitié de rétrotransposons appartenant à trois familles (*RIRE1*, *Kangourou* et *Wallabi*). D'après Piégu *et al.* (2006), ce sont plus de 90 000 copies de ces rétrotransposons qui se sont accumulées dans le génome d'*O. australiensis* au cours des trois millions d'années écoulées depuis sa séparation d'avec le reste de la lignée *Oryza*. Dans l'hypothèse où les génomes de ses plus proches parentes — *O. glaberrima* ($2C = 357$ Mb) et *O. sativa* ($2C = 390$ Mb) — représentent l'état ancestral du génome d'*O. australiensis*, cette espèce a connu au moins un doublement de son génome (247 %), uniquement dû à l'accumulation de rétrotransposons.

Certains rétrotransposons mesurent plus de 20 000 paires de bases et peuvent donc, même avec un faible nombre de copies, influer significativement sur la taille globale du génome. C'est le cas par exemple de la famille *Ogre*[7], un élément Ty3/*gypsy* (Ma-

5. Les *Alu* sont de petits rétrotransposons non-autonomes, les plus nombreux dans le génome humain en terme de nombre de copies. On peut noter également que la fréquence des éléments transposables n'est pas uniforme dans le génome, il y a des différences entre les autosomes et les chromosomes sexuels.

6. Un pseudogène rétroprocessé est le résultat de la transcription inverse d'un ARN messager, l'ADN complémentaire généré étant ensuite inséré dans le génome. Cette nouvelle séquence génomique dépourvue d'introns possède une séquence poly-A à son extrémité 3′ et porte quelque fois les traces d'autres modifications post-transcriptionnelles (Chen *et al.*, 1982 ; Hollis *et al.*, 1982 ; Reilly *et al.*, 1982 ; Graur *et al.*, 1989).

7. Certains éléments *Ogre* présentent également la particularité d'avoir un intron au milieu de la région *gag–pol*. L'épissage des transcrits a été rapporté chez plusieurs rétroéléments comme les rétrovirus,

cas & Neumann, 2007). Chez certaines fabacées, cet élément représente à lui seul une portion importante du génome : environ 5 % des 5 Gb du génome de *Pisum sativum* (Neumann *et al.*, 2003) et, avec 10 000 copies, plus de 38 % du génome de *Vicia pannonica* (1C= 6,75 pg) (Hill *et al.*, 2005 ; Neumann *et al.*, 2006).

D'autres études de même type menées sur l'orge (Vicient *et al.*, 1999 ; Kalendar *et al.*, 2000 ; Schulman & Kalendar, 2005), le blé (Ramakrishna *et al.*, 2002 ; W. Li *et al.*, 2004 ; Chantret *et al.*, 2005), le coton (Hawkins *et al.*, 2008, 2009) ou le riz (Swigoňová *et al.*, 2005 ; Vitte *et al.*, 2007 ; Zuccolo *et al.*, 2007) aboutissent à une conclusion similaire, l'essentiel de la variation de taille de génome est imputable aux éléments transposables de classe I.

2.3 Dynamique des éléments transposables

Des études quantitatives menées sur *Drosophila*, *Homo sapiens*, *Arabidopsis* et *Tetraodon* montrent que les insertions d'éléments transposables sont en général défavorables (Wright *et al.*, 2001 ; Neafsey *et al.*, 2004) et donc susceptibles d'être éliminées par la sélection naturelle. L'efficacité de la sélection E dépend non seulement de la force de la sélection s mais également du nombre d'individus N_e participant à la reproduction[8], selon la formule $E = N_e \times s$. Par conséquent même si la contre-sélection est forte, les transposons peuvent s'accumuler si la population efficace est faible (Petit & Barbadilla, 2009). La prolifération des éléments transposables dans les génomes de plantes pourrait donc être due aussi bien à de faible valeurs s qu'à de faibles valeurs N_e. Les restructurations majeures et l'inflation des génomes observée chez les macro-eucaryotes (les gros organismes) ne seraient donc pas le résultat de processus adaptatifs, mais celui de la levée de la sélection suite à la baisse de la taille de la population effective (Lynch & Conery, 2003 ; Lynch, 2007a)[9].

Hormis la sélection naturelle, l'abondance et l'activité d'un élément transposable dans un génome sont soumises à plusieurs paramètres comme le nombre de copies de l'élément lui-même, les mécanismes de contrôle du génome hôte et certains facteurs environnementaux.

Autorégulation Les éléments autonomes de classe II codent pour une transposase. La présence de cette transposase agit comme un répresseur limitant l'activité de l'élément. C'est le cas par exemple de l'élément *mariner* (Lohe & Hartl, 1996 ; Lohe *et al.*,

les LINE et les éléments de type *Penelope* (Arkhipova *et al.*, 2003 ; M. Tamura *et al.*, 2007). Chez les rétrotransposons, c'est un phénomène plus rare et c'est le seul cas pour lequel la portion épissée ne correspond pas à une région codante. Il s'agit donc d'une situation ressemblant en tous points à l'épissage d'un gène nucléaire classique (Steinbauerová *et al.*, 2008).

8. Il semble également que la taille globale de la population joue un rôle (Veuille *et al.*, 2008).

9. Des données expérimentales sont en contradiction avec cette hypothèse. En forçant l'autofécondation chez *A. thaliana*, la taille de la population efficace doit baisser et les mutations désavantageuses doivent s'accumuler à un rythme plus rapide. Or c'est la situation inverse, une baisse de la charge en éléments transposables, qui est observée par Wright *et al.* (2008). Les mêmes auteurs constatent également une corrélation négative entre la taille du génome et le taux d'autofécondation.

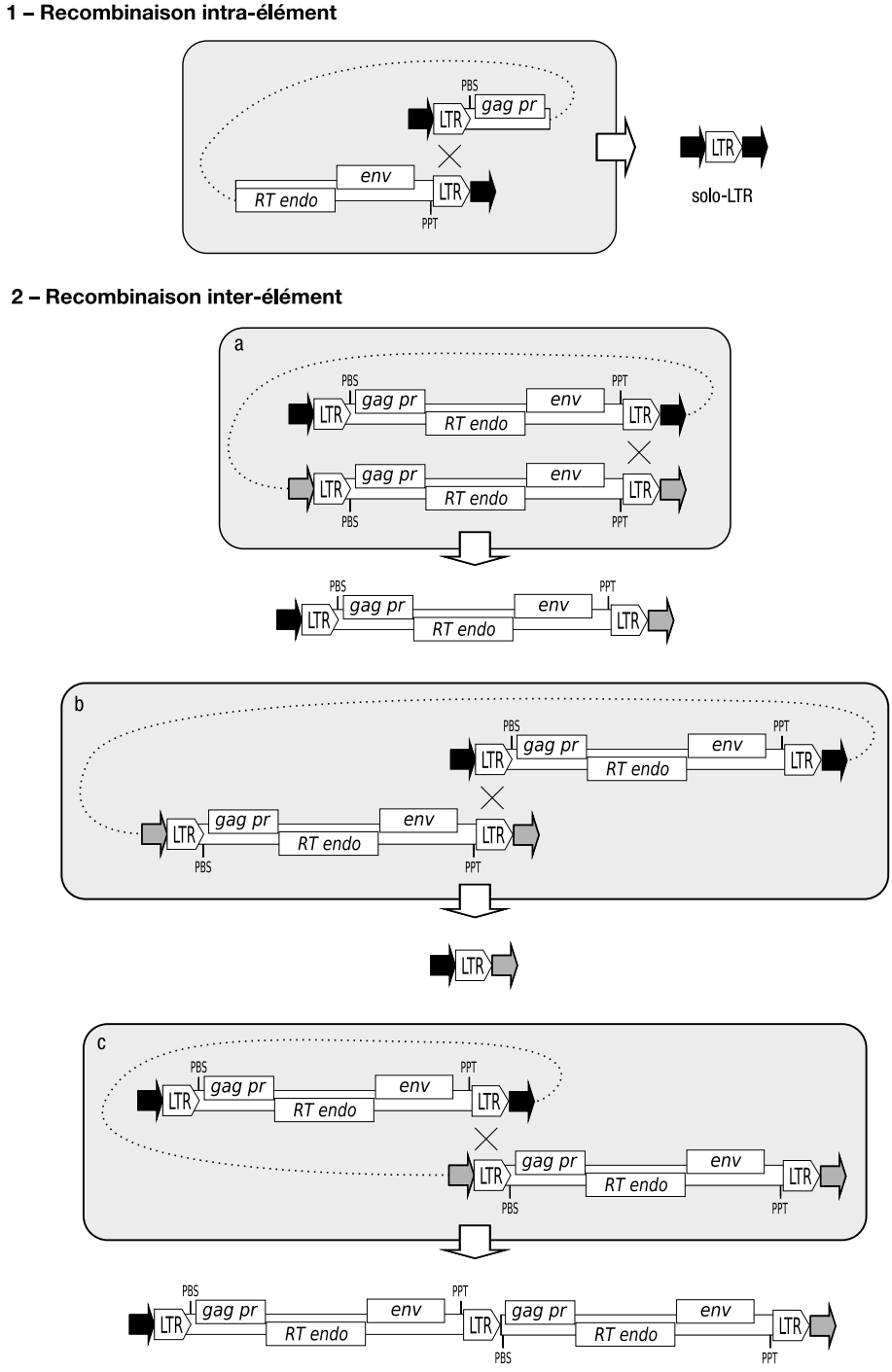

FIG. 2.5 — Différents cas de recombinaisons non-homologues impliquant des rétrotransposons (adapté de Devos *et al.*, 2002). **1 – Recombinaison intra-élément** aboutissant à la formation d'un solo-LTR. La ligne de pointillés représente la boucle formée par la séquence d'ADN. **2 – Recombinaison inter-éléments** entre a) les deux LTR 3′ de deux rétrotransposons adjacents, b) la LTR 5′ et la LTR 3′ de deux rétrotransposons adjacents et c) la LTR 3′ et la LTR 5′ de deux rétrotransposons adjacents. Dans les trois cas, la ligne de pointillés représente la région génique initialement située entre les deux éléments, et perdue suite à l'évènement de recombinaison. Ces mécanismes d'élimination de l'ADN répété jouent un rôle important dans les variations de taille de génome chez les angiospermes (Vitte & Panaud, 2003 ; Ma *et al.*, 2004 ; Gaut *et al.*, 2007).

1996) et de l'élément *P* chez la drosophile (Lemaitre *et al.*, 1993 ; Misra *et al.*, 1993). Des mécanismes similaires existent chez les éléments de classe I, l'action de la *reverse transcriptase* étant chez les rétrovirus sous contrôle de la capside nucléique codée par le gène *gag* (Mougel *et al.*, 2009). Enfin, l'augmentation du nombre d'éléments transposables d'une même famille entraîne une augmentation du nombre de recombinaisons inégales et par conséquent une disparition progressive des copies (Shirasu *et al.*, 2000 ; Devos *et al.*, 2002 ; Wicker *et al.*, 2003 — voir Fig. 1.4 page 20 et Fig. 2.5 page ci-contre).

Régulation par la cellule Les mécanismes épigénétiques — c'est-à-dire l'ensemble des changements d'expression des gènes héritables sans modification de la séquence nucléotidique (C.-t. Wu & Morris, 2001) — peuvent contrôler le nombre et l'activité des éléments transposables *via* la méthylation (Tsukahara *et al.*, 2009) et l'hétérochromatisation de l'ADN (Bestor, 2003), et *via* la dégradation des ARN transcrits (*Non Sens Mediated RNA Decay*)[10] (R. B. Flavell, 1994 ; M. A. Matzke & Matzke, 1998 ; A. J. M. Matzke & Matzke, 1998). Les répétitions terminales contenues dans les éléments transposables en font des cibles naturelles pour la méthyltransférase (Bender, 1998 ; Tompa *et al.*, 2002) et des exemples d'inactivation d'éléments par méthylation ont été décrits chez *A. thaliana* et *Nicotiana tabacum* (Hirochika *et al.*, 2000). De fait, les régions riches en éléments transposables correspondent souvent à de l'hétérochromatine[11], fortement méthylée et compactée, et donc hors de portée du complexe transcriptionnel en conditions normales (Okamoto & Hirochika, 2001).

Facteurs environnementaux et chocs génomiques La méthylation et la compaction de l'hétérochromatine permettent l'inactivation des éléments transposables mais sont des processus coûteux en énergie. En condition de stress, ces mécanismes de contrôle sont partiellement levés, ce qui peut entraîner de nouvelles vagues de transpositions et donc des modifications structurelles du génome (J. Zhang *et al.*, 2009). L'activation des éléments transposables par un stress — blessure, attaque fongique, pollution à l'ozone — a été étudiée chez certains champignons (Ogasawara *et al.*, 2009) et chez les solanacées (Grandbastien *et al.*, 1997 ; Grandbastien, 1998 ; Pourtau *et al.*, 2003 ; Grandbastien *et al.*, 2005). Chez l'orge sauvage *Hordeum spontaneum*, une corrélation positive entre le taux de transposition du rétrotransposon *BARE-1* et le stress hydrique a été mise en évidence par Kalendar *et al.* (2000). Chez le melon, la transposition du rétrotransposon *Reme1* est stimulée par une augmentation des rayonnements ultra-violets (Ramallo *et al.*, 2008) et chez la luzerne, c'est la mise en culture de cellules qui active le rétroélément *MERE1* (Rakocevic *et al.*, 2009). Les facteurs physiques ne sont pas les seuls à engendrer un stress, l'hybridation entraîne elle aussi une modification temporaire des mécanismes de contrôle des éléments transposables (Josefsson *et al.*, 2006).

10. Chez les mammifères, les éléments transposables ne sont pas soumis aux mêmes contrôles dans les cellules somatiques que dans les cellules de la lignée germinale. Les piARN (*Piwi-interacting RNA*), récemment mis en évidence (Filipowicz *et al.*, 2005 ; Obbard *et al.*, 2009), interagissent avec les protéines Piwi et entraînent l'inactivation de rétroéléments endogènes par un mécanisme encore inconnu.

11. Une des conséquences de l'hétérochromatisation est de baisser le taux de recombinaisons et donc de ralentir l'élimination des éléments transposables.

B. Liu & Wendel (2000) rapportent que suite à un croisement entre *Oryza sativa* et l'espèce sauvage *Zizania latifolia*, l'introgression d'ADN en provenance de ce dernier a entraîné des changements de méthylation des rétrotransposons de type *copia* et *gypsy*. Ces changements ont augmenté transitoirement l'activité de transposition de ces éléments avant que les mécanismes de régulation ne les neutralisent à nouveau. Enfin, partant du constat que dans un génome dense l'insertion d'éléments transposables a de forte chance d'être délétère, X. Zhang & Wessler (2004) font l'hypothèse qu'un évènement de polyploïdisation, en introduisant de la redondance fonctionnelle, diminue cette contre-sélection et permette plus d'insertions.

Transferts horizontaux Le passage d'un élément transposable d'un organisme à un autre est un mécanisme indispensable pour expliquer certains résultats de phylogénie moléculaire. De nombreux exemples de transferts horizontaux ont été rapportés chez les animaux, mais seulement quelques-uns chez les plantes (Richardson & Palmer, 2007 ; Fortuné, Roulin & Panaud, 2008, pour revue). C'est le cas par exemple des éléments à LTR de la famille *Rider*. Ces éléments sont constitutivement actifs chez la tomate et sont très similaires aux éléments *Rider* présents chez *A. thaliana*. Les analyses phylogénétiques indiquent qu'il s'agit bien d'un transfert horizontal récent — 1 à 6 millions d'années — qui a conduit à l'introgression et à l'activation de cet élément chez la tomate (Cheng *et al.*, 2009). Les éléments à LTR, structurellement proches des virus, sont susceptibles d'être transférés d'un organisme à un autre plus facilement. À l'inverse, les éléments non-LTR sont considérés comme peu susceptibles de se transmettre horizontalement. Cependant, Novikova *et al.* (2007) ont documenté un possible transfert d'éléments de classe II chez des papillons. Tous les genres étudiés par cette équipe se sont avérés dépourvus de la famille d'éléments *CR1B*, sauf le genre *Maculinea* et deux espèces de la famille Bombycidae : *Oberthueria caeca* et *Bombyx mori*. Étant donné la distance évolutive séparant ces espèces et la faible divergence des séquences de CR1B, l'hypothèse la plus probable est celle d'un transfert horizontal, par le biais d'un vecteur (parasite, bactérie ou virus)[12]. Un cas similaire d'échange de transposons *Mu-like elements* (MULE) a été mis en évidence entre les deux poacées *Oryza* et *Setaria* (Diao *et al.*, 2006). D'autres échanges entre poacées (éléments *RIRE1*) ont été mis en évidence par Roulin *et al.* (2008, 2009). Au cours de ces transferts horizontaux, les éléments transposables peuvent remplir le rôle d'agents — comme les plasmides, les phages et les virus — et véhiculer d'un organisme à l'autre des portions d'ADN génomique. Cette catégorie d'ADN, désignée sous le terme « mobilome » et dont l'exploration ne fait que commencer, joue un rôle moteur dans l'évolution de la biosphère (Frost *et al.*, 2005 ; Siefert, 2009).

12. Les mécanismes de transfert sont probablement les mêmes chez les angiospermes.

2.4 Changements phénotypiques liées aux éléments transposables

L'accumulation d'éléments transposables dans les génomes n'est pas sans conséquences. Il est clair que les éléments transposables, de par leur nombre et leur mobilité, peuvent entraîner des changements plus ou moins importants au niveau du transcriptome (Weil & Wessler, 1990). L'insertion d'un élément à proximité d'un gène, ou au sein même du gène, peut avoir des conséquences importantes voir létales. Les modifications les plus graves étant contre-sélectionnées, seules les modifications positives ou faiblement négatives nous sont visibles.

Dans des lignées de cellules humaines cancéreuses, il a été montré que la levée de la méthylation, ou hypométhylation, des éléments LINE-1 et *Alu* — éléments non-LTR constituant plus de 30 % du génome humain — conduit à une forte instabilité génomique (Daskalos *et al.*, 2008).

Chez l'humain et la souris, il existe des sites de terminaison alternatifs (sites poly-A) associés à des éléments transposables. Les éléments transposables peuvent donc modifier la région 3' de certains gènes en introduisant des sites poly-A alternatifs, et donc générer de la diversité (J. Y. Lee *et al.*, 2008). Certains éléments — *Tgm-Express*, Pack-MULE, hélitrons et quelques rétroéléments — sont connus pour être capables de transporter de l'ADN génomique (Hollister & Gaut, 2007). Ils peuvent non seulement modifier les gènes existants mais également introduire de nouveaux gènes et enrichir le protéome de l'organisme hôte.

Chez les primates, la famille de gène IRG (*Immunity related GTPases*) joue un rôle important dans la défense contre les bactéries intracellulaires. Un membre de cette famille, le gène IRGM a été identifié comme un des nombreux facteurs génétiques liés à la maladie de Crohn[13] (Parkes *et al.*, 2007). L'étude de la région environnant ce gène, menée chez plusieurs espèces — incluant le chien, la souris et plusieurs primates — révèle une histoire complexe et liée à des mouvements d'éléments transposables (Bekpen *et al.*, 2009). Le scénario est le suivant. Après la divergence Strepsirrhiniens–Haplorrhiniens, le gène IRGM devient simple-copie. Il y a 40 millions d'années, chez l'ancêtre des Haplorrhiniens, ce gène IRGM unique est « éteint » par insertion d'un élément *Alu* dans le second exon. Il y a 24 à 18 millions d'années, trois évènements successifs vont réactiver le gène dans la lignée des Hominoïdes (gibbon, orang-outang, gorille, homme, chimpanzé). Un rétrovirus ERV9 s'intègre en amont du gène et agit comme un promoteur, la mutation d'un nucléotide crée un nouveau codon ATG (codon d'initiation) après l'élément *Alu*, puis la mutation d'un codon stop permet la traduction complète. Chez les gibbons et les orangs-outans, des copies fonctionnelles peuvent côtoyer des copies non-fonctionnelles. Chez les Hominidés par contre, la copie fonctionnelle est fixée et le gène s'exprime normalement. Pour Bekpen *et al.* (2009), il s'agit peut être du premier cas documenté d'extinction puis de réveil d'un gène sous l'influence d'insertions d'éléments transposables.

13. La maladie de Crohn est une maladie inflammatoire chronique de l'ensemble du tube digestif touchant environ 1 personne sur 1 000.

D'autres exemples de modification de l'expression génique par des éléments transposables ont été rapportés chez des angiospermes. Au sein du genre *Petunia*, la production d'anthocyanines est sous contrôle du gène anthocyanin2 (*an2*). Ce gène est donc un des déterminants de la couleur des fleurs. Or, chez la sous-espèce à fleurs blanches *Petunia axillaris*, le gène *an2* a été inactivé suite à l'insertion puis l'excision d'un transposon (Quattrocchio *et al.*, 1999). Le gène *Hf1* intervenant également dans la couleur des pétales est lui aussi modifié par des insertions d'éléments transposables (Matsubara *et al.*, 2005). La couleur et la forme des fleurs étant des paramètres importants pour attirer des polinisateurs (oiseaux, insectes), une altération de ces caractères peut avoir des conséquences importantes, pouvant conduire à un isolement génétique et, à terme, à une spéciation sympatrique[14]. Chez la vigne, *Vitis vinifera*, c'est également l'accumulation d'anthocyanines qui est responsable de la couleur du raisin noir et c'est l'insertion d'un rétrotransposon en amont d'un gène régulant la production d'anthocyanines qui aurait permis l'apparition de cultivars blancs (Kobayashi *et al.*, 2004). Chez le pois, *Pisum sativum*, le phénotype « graines ridées » est du à l'inactivation par un transposon d'un gène codant pour une enzyme de structure et de cohésion de l'amidon (Bhattacharyya *et al.*, 1990). Enfin, chez certaines variétés de tomate, le rétrotransposon *Rider* entraîne un déplacement et une surexpression du gène *Sun*. La forme du fruit et le développement des graines s'en trouvent profondément modifiés (Xiao *et al.*, 2008).

2.5 Domestication des éléments transposables

Le matériel génétique nouveau que constituent les éléments transposables peut être réutilisé et devenir un élément constitutif du génome. Ce processus est baptisé « domestication » et plusieurs exemples ont pu être mis évidence (Xing *et al.*, 2006 ; X. H.-F. Zhang & Chasin, 2006 ; González *et al.*, 2009). Chez les métazoaires, un travail de modélisation montre que les éléments transposables ont pu jouer un rôle important dans l'apparition de la différenciation cellulaire, et plus précisément dans l'apparition de la lignée germinale (Johnson, 2008 ; Filatov *et al.*, 2009). Chez les mammifères, des éléments SINE influencent l'expression de deux gènes-clés dans le développement du néocortex mammalien (Sasaki *et al.*, 2008), des rétrotransposases d'éléments de type *pogo* ont été recrutées comme protéines centromériques (Casola *et al.*, 2008), le gène *Rag* dérive de la transposase d'une ancienne famille de transposons appelée *Transib* (Gent *et al.*, 1996 ; Kapitonov & Jurka, 2005) et les gènes *Mart* dérivent de l'élément *Sushi*-like présent uniquement chez les mammifères (Brandt *et al.*, 2005).

14. Dans le cas de *Petunia axillaris*, il semble cependant que l'inactivation du gène *an2* soit postérieure à l'isolement génétique de la sous-espèce. Toutefois, un isolement génétique induit par des éléments transposables peut se dérouler de la façon suivante. Les remaniements chromosomiques dus aux transposons peuvent empêcher l'appariement des chromosomes chez les hybrides, et la présence de transposons peut modifier l'expression de gènes impliqués dans le développement des individus et modifier la période de maturité sexuelle. Dans les deux cas, un isolement reproductif se met en place et la spéciation peut se faire (Anxolabéhère *et al.*, 2007).

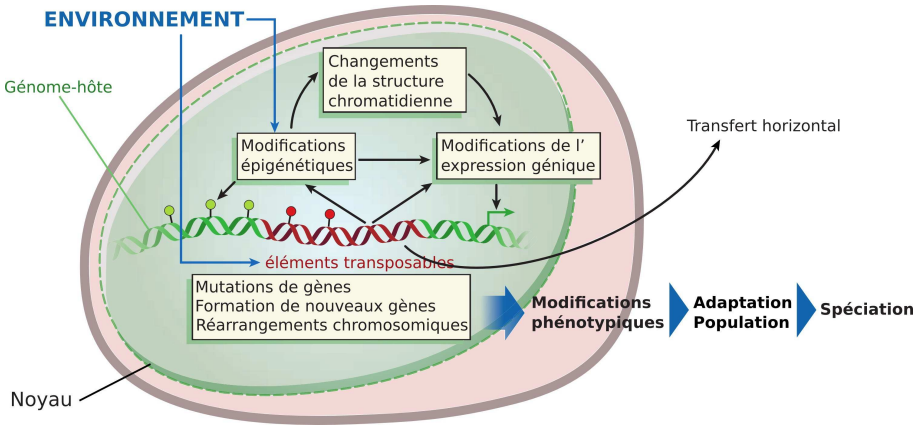

FIG. 2.6 — Éléments transposables et interactions génome–environnement. Sous l'action de l'environnement, les éléments transposables peuvent s'activer et entraîner des modifications structurales et fonctionnelles du génome-hôte. Ces mutations peuvent entraîner des changements phénotypiques, un changement dans la valeur adaptative de l'individu et, éventuellement un phénomène de spéciation. Les éléments transposables peuvent également être transférés horizontalement et accéder ainsi à de nouveaux territoires. L'hyperméthylation est un des principaux mécanismes limitant l'activité des éléments transposables. D'après Biémont & Vieira (2006).

Selon Feschotte (2008), chez *Homo sapiens* 10 000 fragments d'éléments transposables semblent avoir été conservés depuis l'émergence du groupe des mammifères placentaires au début du Crétacé (145,5–99,6 millions d'années). Les régions codantes du génome humain sont composées à environ 4 % d'éléments transposables[15] (Nekrutenko & Li, 2001). Sorek *et al.* (2002) ont montrés que les exons portant des éléments transposables de type *Alu* sont transcrits, mais pour la plupart éliminés par épissage alternatif. Les rétrovirus HERV (*Human endogenous retroviruses*) peuvent être associés à des maladies ou à des désordres cellulaires graves. Cependant, certains de ces virus ont été domestiqués par la cellule. C'est le cas du gène codant la syncytine, une protéine indispensable à la formation du placenta et dérivant du gène *env* d'un virus endogène (Mi *et al.*, 2000 ; Frendo *et al.*, 2003 ; Dupressoir *et al.*, 2009).

Les éléments transposables contribuent donc au transcriptome, mais est-il possible de trouver leur trace dans le protéome ? Une première étude faite sur la *Protein Data Bank*[16] montre que seules 3 entrées sur les 3 000 que comptait la base à cette époque contiennent des séquences d'éléments transposables (Gotea & Makałowski, 2006). Une autre étude confirme l'épissage alternatif de 88 % des exons porteurs d'éléments transposables (M. Wu *et al.*, 2007). Les 12 % restant (30 exons sur 248) sont constitutifs —

15. Si théoriquement, n'importe quel portion d'élément transposable peut-être recrutée, il apparaît que les *transposases* — portées par les éléments de classe II — sont les plus couramment domestiquées (43 des 47 transposases connus chez l'humain), alors que ces mêmes éléments ne représentent que 7 % des éléments transposables présent dans le génome humain (Feschotte, 2008). Par exemple, le fragment de transposon *Hsmar1* intégré dans la protéine SETMAR a toujours son ancienne fonction de fixation de la transposase (Cordaux *et al.*, 2006 ; Jordan, 2006).

16. `http://www.rcsb.org/pdb/home/home.do`

c'est-à-dire présents chez tous les transcrits du gène — et contiennent des éléments transposables récents et non-dégénérés. Chez *Arabidopsis*, deux cas de gènes recrutés à partir de séquences de rétrotransposases ont été rapportés (Hudson *et al.*, 2003).

En résumé, les éléments transposables jouent un rôle majeur dans l'évolution des organismes. En modifiant en profondeur la structure des chromosomes, en agissant sur le fonctionnement et l'expression des gènes, en introduisant des possibilités de recombinaisons, de transferts et de réutilisation de matériel génétique, ils créent de la diversité et peuvent à terme déclencher ou accompagner des processus d'adaptation et de spéciation (voir Fig. 2.6 page précédente).

3

Le genre *Lupinus* (Tourn.) L. 1753

« Qu'il n'est pas vrai que les espèces soient aussi anciennes que la nature, et qu'elles aient toutes existé aussi anciennement les unes que les autres ; mais qu'il l'est qu'elles se sont formées successivement, qu'elles n'ont qu'une constance relative et qu'elles ne sont invariables que temporairement. »

Jean-Baptiste Pierre Antoine DE MONET, Chevalier de LAMARCK (1744–1829)[1]

DANS ce troisième chapitre, nous nous attacherons à décrire le genre *Lupinus*, sa position dans l'arbre du vivant et sa répartition géographique naturelle. Nous dresserons un panorama des connaissances actuelles (systématique) et des multiples intérêts du genre *Lupinus* (écologie, agriculture, alimentation, santé).

Le genre *Lupinus* a été initialement décrit par Joseph Pitton de Tournefort[2] dans son ouvrage *Institutiones rei herbariæ* (1700, pages 392 et 393). Il crée le niveau taxonomique du *genre* — qu'il pense comme un groupe « naturel » —, et regroupe 17 espèces sous l'épithète *Lupinus*. En 1753, Linné crée les règles de sa nomenclature binomiale, entraînant une nouvelle description du genre *Lupinus* (Kasprzak *et al.*, 2006).

1. Entête du troisième chapitre de la première partie de *Philosophie zoologique* (1809).

2. Joseph Pitton de Tournefort, botaniste français né le 5 juin 1656 à Aix-en-Provence et décédé le 28 décembre 1708 à Paris (Becker *et al.*, 1957, pour revue). Il se prend de passion très tôt pour la botanique et constitue plusieurs herbiers — Savoie, Dauphiné et Pyrénées — qui lui valent d'obtenir le poste de suppléant du surintendant au Jardin du roi en 1693. Sur ordre royal, il repart herboriser dans les Pyrénées. Entre 1700 et 1702, il voyage dans les îles grecques et visite ensuite Constantinople, les côtes de la mer Noire, l'Arménie et la Géorgie. Le récit de son périple, fera l'objet d'un livre : *Relation d'un voyage au Levant* (1717). Il propose dans ses *Élémens de botanique, ou méthode pour connoître les plantes* (1694), un système de classification didactique et efficace regroupant les plantes suivant la forme de leurs corolles et de leurs fruits, mais, plus important encore, il forge les concepts de *famille* et de *genre*, et prépare ainsi le terrain pour Linné. Ce dernier lui dédia le genre *Tournefortia* (Boraginaceae).

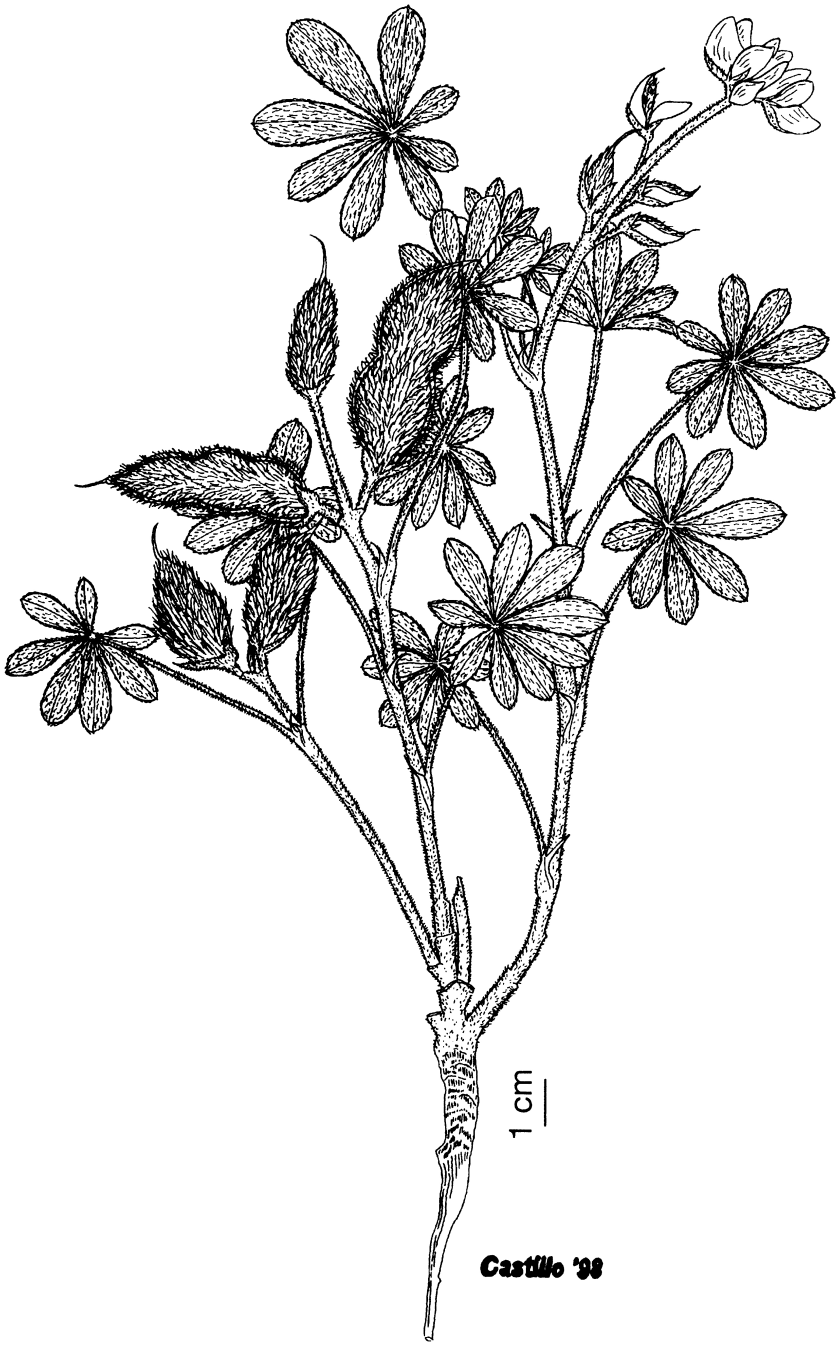

FIG. 3.1 — Anatomie générale de *Lupinus cosentinii* (Scabrispermae) dessiné par X. Castillo et tiré de Castroviejo & Pascual (1999).

3.1 Position systématique

La taxonomie traditionnelle place les lupins dans le règne Plantae, sous-règne Tracheobionta, division Magnoliophyta, classe Magnoliopsida, sous-classe Rosidae, ordre des Fabales, famille des Fabaceae, tribu des Genisteae (Adanson) Bentham[3]. Les progrès de la reconstruction des relations phylogénétiques des organismes et l'accumulation des données moléculaires au cours des trente dernières années ont conduit à une refonte de ce système de classification.

D'après le travail de compilation de l'*Angiosperm Phylogeny Group* (2003), le chemin menant au genre *Lupinus* est le suivant : Angiospermes, Eudicotylédones, Rosidées, Eurosidées I, Fabales[4], Fabaceae[5], Faboideae[6], Genisteae (voir Fig. 3.2 page suivante).

Les eudicotylédones comprennent deux clades principaux, le clade des astéridées (80 000 espèces) et le clade des rosidées (70 000 espèces selon l'*Angiosperm Phylogeny Group*, 2003). Ce dernier formant un groupe phylogénétiquement mal résolu (Doyle & Luckow, 2003 ; H. Wang *et al.*, 2009). Les rosidées seraient apparues il y a 117 à 108 millions d'années selon Wikström *et al.* (2001), 115 à 112 millions d'années selon Davis *et al.* (2005) et 115 à 93 millions d'années selon H. Wang *et al.* (2009) et leur émergence serait liée à l'essor des forêts d'angiospermes et à la diversification de plusieurs groupes d'insectes (Farrell, 1998 ; Wilf *et al.*, 2000 ; Moreau *et al.*, 2006 ; Lengyel *et al.*, 2009). Les données génétiques s'accumulent rapidement pour ce groupe puisque quatre des cinq premiers génomes d'angiospermes publiés appartiennent aux rosidées : *Arabidopsis* (Brassicaceae), *Carica* (Caricaceae), *Populus* (Salicaceae) et *Vitis* (Vitaceae)[7]. Deux autres génomes, *Manihot* et *Ricinus* (Euphorbiaceae), sont également en cours de séquençage.

Les rosidées se subdivisent en deux clades, eurosidées I (ex : *Populus* et *Cucumis*) et eurosidées II (ex : *Arabidopsis* et *Gossypium*). Les eurosidées I regroupent la quasi totalité des espèces capables de s'associer à des bactéries fixatrices d'azote, et notamment les clades fabales, fagales, rosales et cucurbitales (voir Fig. 3.3 page 41).

Les lupins appartiennent à l'ordre des fabales, famille des fabacées (aussi appelées légumineuses) et sous-famille des faboïdées. Les fabacées constituent la troisième plus grande famille au sein des angiospermes avec 19 400 espèces réparties en cinq sous-familles — Cercideae, Detarieae, Caesalpinioideae, Mimosoideae et Faboideae — et 730 genres à travers le globe (Lewis *et al.*, 2005, cité par Jansen *et al.*, 2008). Il s'agit éga-

3. Pour les noms d'espèces et pour les taxons de rangs supérieurs, je me suis basé sur la nomenclature publiée par l'*International Association for Plant Taxonomy* en 1999 (Code de Saint-Louis — version française sur `http://www.tela-botanica.org/page:code_botanique_st_louis`) et en 2006 (Code de Vienne — `http://www.ibot.sav.sk/karolx/kod/0000Viennatitle.htm`).

4. Auparavant, l'ordre des Fabales ne comprenait qu'une seule famille : les Fabaceae. Les progrès récents de la méthode phylogénétique, alliés à l'accumulation de données moléculaires ont conduit à lui rattacher les familles Polygalaceae, Quillajaceae et Surianaceae.

5. Anciennement Leguminosae, cette famille inclut désormais les Caesalpiniaeae et les Misosaceae, en sus des Faboideae.

6. Les Faboideae, aussi appelées Papilionoideae, regroupent 13 855 des 19 400 espèces de Fabaceae et sont caractérisés par leurs fleurs en forme de papillon.

7. Le cinquième étant celui d'*Oryza sativa* (Monocotylédones).

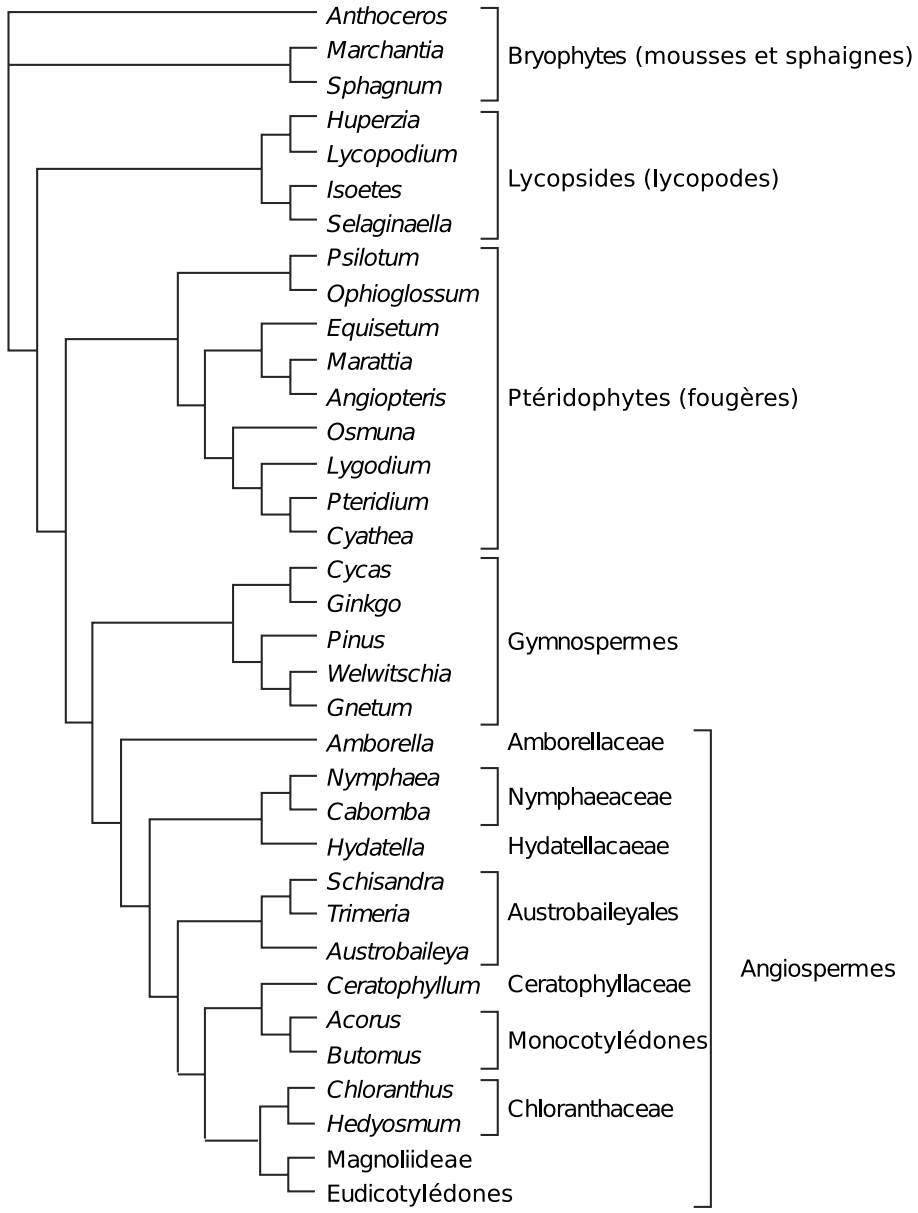

FIG. 3.2 — Phylogénie des embryophytes basée sur les données de Pryer *et al.* (2001) ; Soltis *et al.* (2005) et publiée par Frohlich & Chase (2007). Les relations au sein des angiospermes ne sont pas complètement clarifiées et une autre topologie — plus proche de celle de l'*Angiosperm Phylogeny Group* (2003) — a été proposée par Saarela *et al.* (2007). Les eudicotylédones comprennent deux clades principaux, les astéridées et les rosidées (les lupins appartiennent à ce dernier).

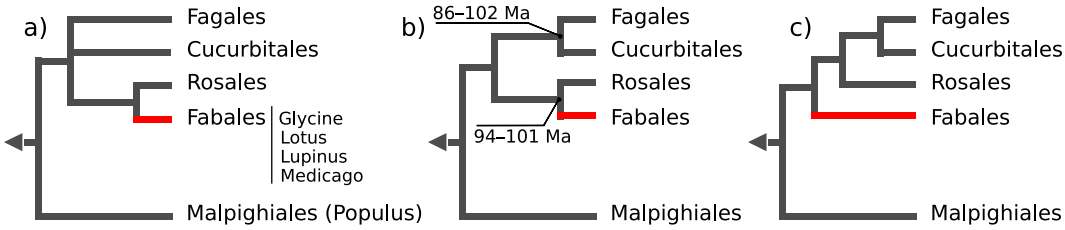

FIG. 3.3 — Incertitude sur la position phylogénétique des Fabales. Les différentes topologies sont proposées par a) l'*Angiosperm Phylogeny Website* (Stevens, 2001), b) Magallón & Castillo (2009), et c) H. Wang *et al.* (2009).

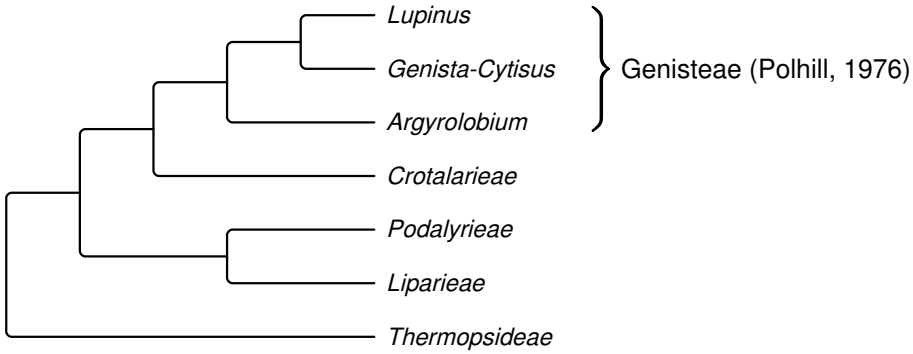

FIG. 3.4 — Position phylogénétique du genre *Lupinus* à l'intérieur de la tribu des Genisteae *sensu* Polhill (1976), basée sur les séquences de l'ITS, du *rbc*L et du *trn*L-*trn*F (Käss & Wink, 1997a ; Aïnouche & Bayer, 1999 ; Aïnouche *et al.*, 2003).

lement de la plus importante famille d'arbres dans les forêts tropicales d'Amérique du Sud et d'Afrique, et de la deuxième dans les forêts du Sud-Est asiatique derrière les dipterocarpacées, une famille de rosidées (Gentry, 1988). Très utilisées dans l'agriculture pour leur graines riches en protéines, les fabacées ne sont devancées que par les poacées en terme d'importance économique et écologique.

Enfin, les génistées appartiennent au groupe des faboïdées (voir Fig. 3.4). Elles produisent toutes des alcaloïdes de type quinolizidine, et seraient originaires de l'Ancien Monde (Crisp *et al.*, 2003).

3.2 Distribution géographique naturelle

Le genre *Lupinus* s'étend sur quatre continents : Afrique, Europe, Amérique du Nord et Amérique du Sud (répartition qualifiée d'amphiatlantique). Si le nombre de taxons décrits est faible pour l'Ancien Monde — 17 à 18 espèces et sous-espèces — le Nouveau Monde a été le siège d'une intense radiation et compte aujourd'hui plusieurs centaines d'espèces, le compte variant selon les auteurs : de 200 à 500 espèces pour Dunn (1984), environ 275 pour Hughes & Eastwood (2006) et plus de 300 pour

Kasprzak *et al.* (2006). De fait, le nombre de dénominations botaniques publiées pour le genre *Lupinus* est très important. Le site *ILDIS World Database of Legumes*[8] donne une liste de 639 noms valides. Le site *Integrated Taxonomic Information System*[9] donne quant à lui une liste de 944 noms, dont 366 noms valides. La taxonomie des différentes espèces de lupins est basée principalement sur cinq caractères — annuel ou pérenne, morphologie des feuilles et des cotylédons, microstructure des téguments des graines et nombre chromosomique — et a abouti à la reconnaissance de plusieurs groupes stables.

3.2.1 Lupins du Nouveau Monde

Les lupins du Nouveau Monde couvrent un large spectre d'habitats, du climat sub-arctique de l'Alaska à celui de la forêt tropicale sud-américaine, des plaines côtières aux hauts plateaux andins. Ils présentent également une grande variété de traits d'histoire de vie : annuels ou pérennes, à port herbacé ou buissonnant, auto- ou allogames. Cette diversité morphologique, biologique et écologique a posé de grandes difficultés taxonomiques, qui ont pu être en partie résolues par l'arrivée de la phylogénie moléculaire. Deux grands groupes monophylétiques remarquables ont été mis en évidence par Aïnouche & Bayer (1999) et Aïnouche *et al.* (2004), le groupe des lupins est-américains et le groupe des lupins ouest-américains. Plus récemment, un échantillonnage plus large dans le Nouveau Monde a permis de confirmer les deux groupes précédents et de distinguer un troisième groupe monophylétique de lupins unifoliolés endémiques du Sud-Est des États-Unis (voir Fig. 3.5 page suivante).

Lupins de l'Est D'après Monteiro & Gibbs (1986), une trentaine d'espèces de lupins sont considérées comme unifoliolées ou alliées aux unifoliolées (c'est-à-dire présentant des feuilles unifoliolées au début de leur développement, ou comme *L. paraguarensis*, présentant sur un même pied les deux types de feuilles). Ces espèces, toutes pérennes, et présentant des nombres chromosomiques allant de $2n = 32$ à $2n = 36$ — exceptionnellement $2n = 52$ (Maciel & Schifino-Wittmann, 2002 ; Conterato & Schifino-Wittmann, 2006) — sont réparties de la façon suivante : vingt-sept espèces dans le Sud-Est de l'Amérique du Sud — hautes terres de l'Est brésilien, du Paraguay, de l'Uruguay et de l'Argentine (Planchuelo & Dunn, 1984) —, deux espèces au Sud des États-Unis (*L. texensis* et *L. harvardii*) et quatre espèces dans le Sud-Est de l'Amérique du Nord (Floride et États voisins).

Selon Dunn (1971), ces quatre espèces — *L. cumulicola*, *L. diffusus*, *L. villosus* et *L. westianus* — auraient divergé récemment à partir d'un transfert longue distance depuis l'Est de l'Amérique du Sud. Les données moléculaires viennent contredire cette hypothèse en montrant que les lupins unifoliolés du Sud-Est des États-Unis ne sont pas apparentés aux lupins unifoliolés d'Amérique du Sud. Ce caractère serait donc apparu deux fois dans l'histoire évolutive des lupins (Hughes & Eastwood, 2006 ; Eastwood *et al.*, 2008).

8. `http://www.ildis.org/LegumeWeb/6.00/names/npall/npall_427.shtml`
9. `http://itis.gov/`

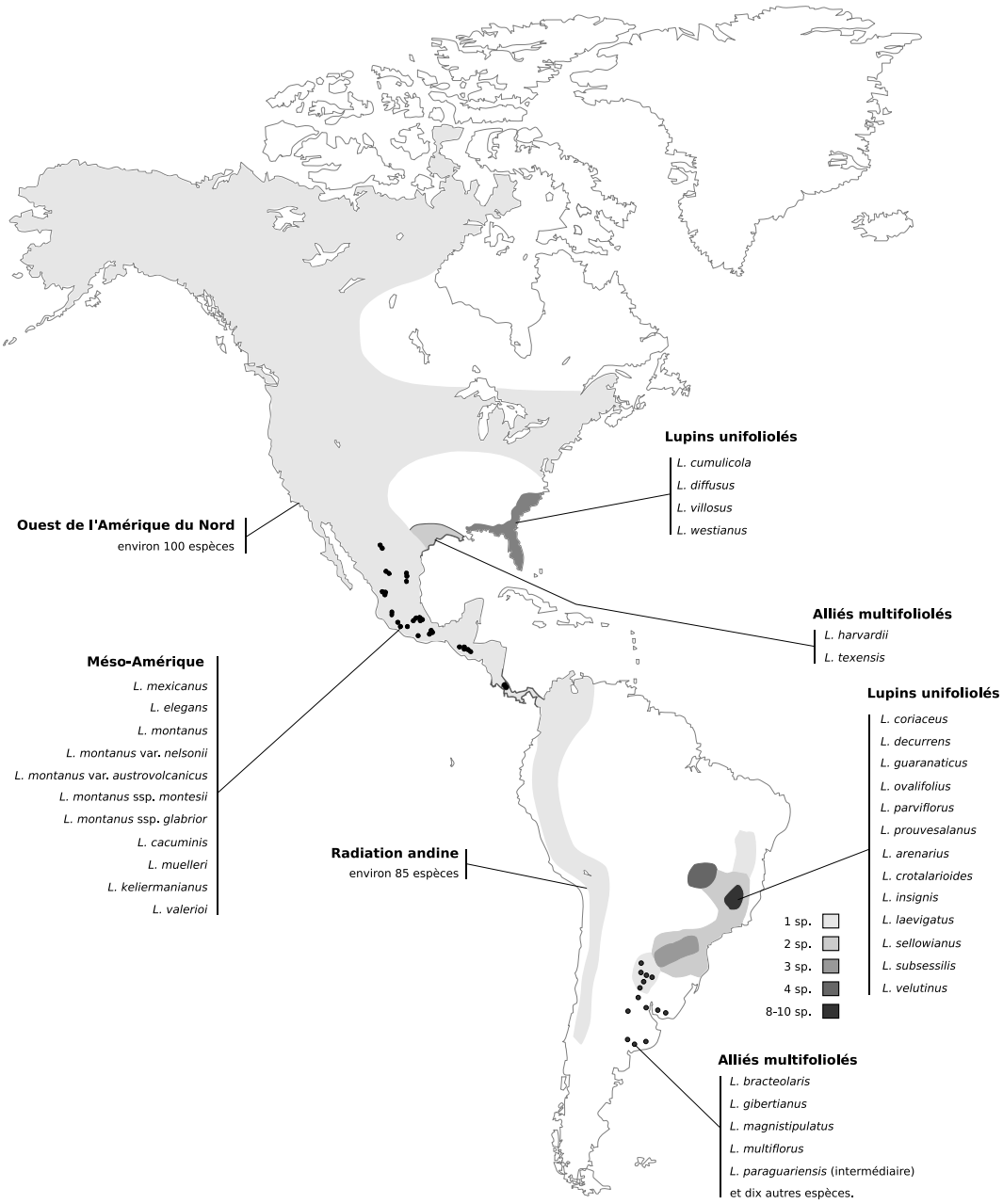

FIG. 3.5 — Distribution des lupins du Nouveau Monde. Les lupins couvrent un large spectre d'habitats et un grand nombre d'espèces ont été décrites. Deux grands clades se détachent (1) le clade des lupins à feuilles unifoliolées du Sud-Est des États-Unis (Dunn, 1971) et de l'Est de l'Amérique du Sud, avec leurs alliés à feuilles multifoliolées (Planchuelo & Dunn, 1984) et (2) le clade des lupins ouest-américains, très important en nombre d'espèces (Planchuelo, 1994). Ce dernier est composé d'un groupe d'une centaine d'espèces colonisant l'Amérique du Nord (Dunn & Gillett, 1966), d'un groupe méso-américain (Dunn & Harmon, 1977) et d'un groupe andin d'environ 85 espèces (Hughes & Eastwood, 2006). La diversification des lupins andins est liée au processus géologique de surrection des Andes. Il s'agit d'une radiation récente et très rapide, comparable à celle des cichlidés des grands lacs africains (Hughes & Eastwood, 2006).

Lupins de l'Ouest Les espèces de l'Ouest se répartissent en trois grands centres de diversité : l'Amérique du Nord avec une centaine d'espèces, les Andes avec environ 85 espèces (Hughes & Eastwood, 2006) et l'Amérique centrale avec un nombre plus réduit d'espèces (Dunn & Harmon, 1977). Ces espèces présentent une grande diversité morphologique (port herbacé, buissonnant ou arbustif). À l'inverse, elles présentent un nombre chromosomique très stable à $2n = 48$ (plus rarement $2n = 96$). Les lupins de l'Ouest de l'Amérique du Sud colonisent principalement la cordillère des Andes. Cette occupation des hautes terres se prolonge vers le Nord à travers l'Amérique centrale puis à travers le Mexique (Sierra Madre Occidentale et Orientale), les États-Unis et le Canada, principalement dans les régions montagneuses situées à l'Ouest du 100^{e} méridien (Sierra Nevada, chaînes côtières du Pacifique, Montagnes Rocheuses, chaîne des Cascades et chaîne d'Alaska). La phylogénie des lupins de l'Ouest, particulièrement complexe, a fait l'objet ces dernières années d'un travail important (Hughes & Eastwood, 2006 ; Ch. S. Drummond, 2008 ; Eastwood *et al.*, 2008).

3.2.2 Lupins de l'Ancien Monde

Tous les lupins de l'Ancien Monde sont annuels, se reproduisant majoritairement par autofécondation, avec un port herbacé et des feuilles digitées (Gladstones, 1974 ; Heyn & Herrnstadt, 1977 ; Plitmann & Heyn, 1984 ; Gladstones, 1998)[10]. Les espèces de l'Ancien Monde se divisent en deux groupes sur la base de la microstructure des téguments de leurs graines : les lupins à graines rugueuses (*Scabrispermae*) et les lupins à graines lisses.

Les lupins à graines rugueuses Les lupins à graines rugueuses ont une distribution principalement africaine (voir Fig. 3.6 page ci-contre). On dénombre neuf espèces et sous-espèces, présentant des nombres chromosomiques allant de $2n = 32$ à $2n = 42$. Certaines de ces espèces décrites récemment, ont une distribution très limitée et un statut taxonomique incertain. C'est le cas de *Lupinus anatolicus*, décrit par Święcicki *et al.* (1996), suite à une mission de collecte effectuée en 1977 dans le Sud-Ouest de la Turquie, sur les collines situées entre les villes de Selçuk et de Kuşadasi. C'est le cas également de *Lupinus atlanticus*, décrit par Gladstones (1974) et présent uniquement au Maroc ou encore de *Lupinus pilosus* var. *tassilicus* (ou *L. tassilicus*) présent uniquement sur les plateaux gréseux du Sahara central.

Les lupins à graines lisses Les lupins à graines lisses ont une distribution principalement méditerranéenne (voir Fig. 3.7 page 46). On dénombre neuf espèces et sous-espèces, présentant des nombres chromosomiques allant de $2n = 40$ à $2n = 52$. Comme pour les lupins africains, la synonymie est importante et le statut de certains

10. John S. Gladstones est un scientifique australien, créateur en 1967 d'*Uniwhite*, la première variété de lupin adaptée à une culture à grande échelle. Il est l'auteur d'un travail remarquable de description du genre *Lupinus* et de collecte de données écologiques, physiologiques et agronomiques.

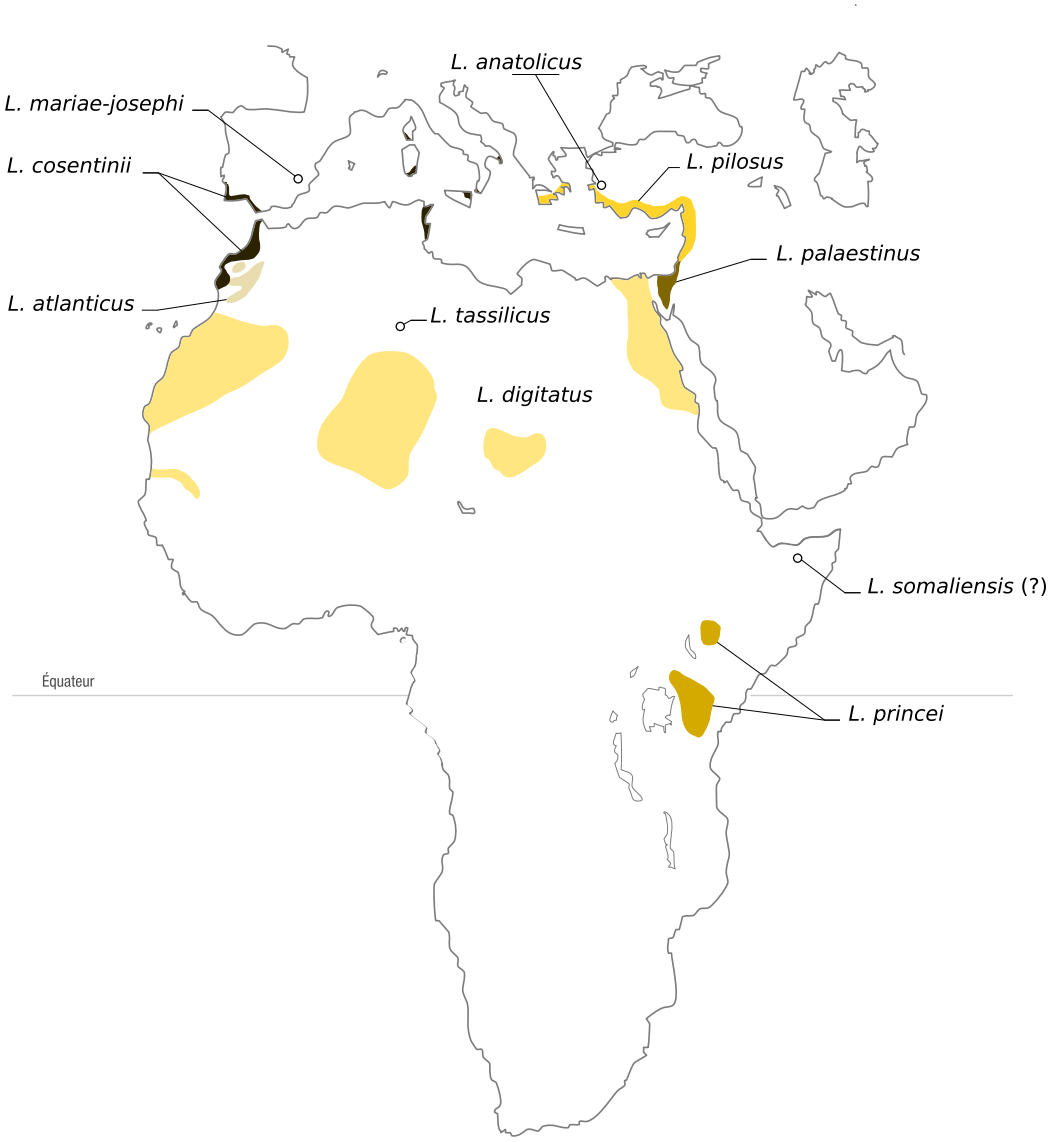

FIG. 3.6 — Distribution des lupins à graines rugueuses de l'Ancien Monde d'après Gladstones (1998). *Lupinus princei* vit en région équatoriale, sur les hauts-plateaux d'Afrique de l'Est. *Lupinus atlanticus* est présent uniquement au Maroc tandis que *L. cosentinii* est présent sur les côtes marocaines, portugaises et tunisiennes ainsi qu'en Corse, en Sardaigne, en Sicile et dans le sud des Pouilles italiennes. *Lupinus pilosus* et *L. anatolicus* occupent la côte Est de la Méditerranée. *Lupinus palaestinus* est endémique de la région englobant le désert du Néguev et le désert du Sinaï, tandis que *L. digitatus*, typique de la vallée du Nil, serait également présent au Sahara central et sur la côte Atlantique. *Lupinus pilosus* var. *tassilicus* est présent uniquement sur les plateaux gréseux du n'Ajjer algérien. *Lupinus mariae-josephi* est une espèce découverte en Espagne dans la région de Valence (Pascual, 2004) et présentant des caractères intermédiaires entre les lupins à graines rugueuses et les lupins à graines lisses (arbitrairement placée avec les Scabrispermae sur cette figure). Enfin, *L. somaliensis* n'est connu qu'à travers un seul spécimen, collecté par Edith Cole & Lort Phillips et décrit par John G. Baker en 1895. Selon Gladstones (1974), son statut d'espèce ne fait pas de doute.

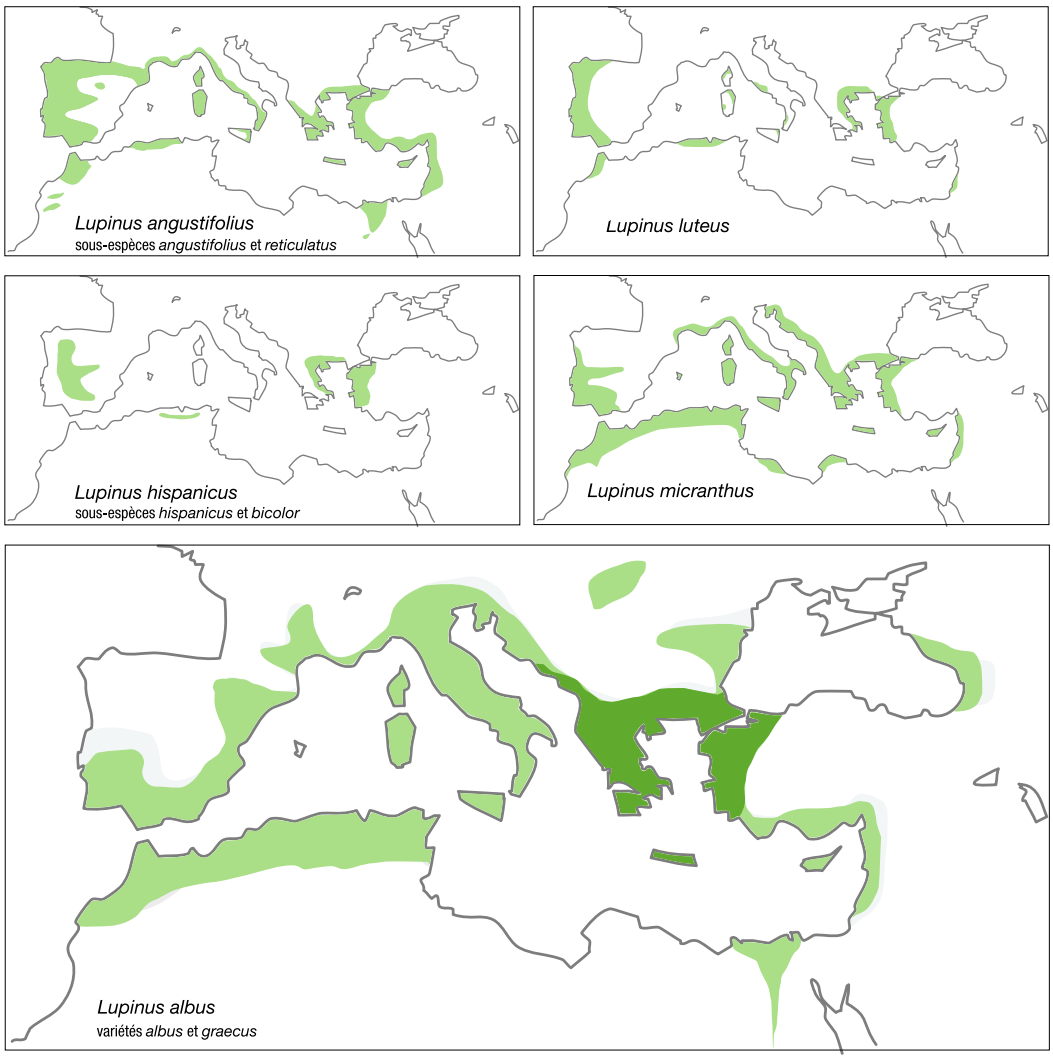

FIG. 3.7 — Distribution des lupins à graines lisses de l'Ancien Monde d'après Gladstones (1998). Une population de *L. angustifolius* est présente en Bretagne Sud, sur l'île d'Hœdic, mais d'après Gladstones, l'extension naturelle de cette espèce sur la façade atlantique ne dépasse pas les Pyrénées. *Lupinus luteus* est restreinte aux sols sableux tandis que *L. micranthus* a une large tolérance édaphique. L'espèce *L. albus* est présente sur une grande partie du pourtour méditerranéen et a été décrite sous plusieurs noms différents : *L. graecus* en Grèce, *L. termis* en Égypte, *L. vavilovii* et *L. jugoslavicus* dans les Balkans (Przyborowski & Weeden, 2001).

taxons est incertain. C'est le cas de *Lupinus vavilovii*[11], décrit par Atabekova & Maissurjan, suite à une mission de collecte dans la région des Balkans. Przyborowski & Weeden (2001) ont proposé que cette dénomination soit déclassée en synonyme de la sous-espèce *L. albus* L. var. *graecus* — elle même encore souvent appelée *L. graecus* par certains auteurs (Naganowska *et al.*, 2003).

En 2004, Higinio Pascual a décrit *Lupinus mariae-josephi*, une espèce nouvelle endémique de la région de Valence en Espagne. Cette espèce présente des caractères morphologiques originaux ne permettant pas de la rattacher clairement à l'un ou l'autre des groupes reconnus de lupins. Cette nouvelle espèce, ainsi que les populations au statut taxonomique incertain précédemment citées, ont fait l'objet d'une attention particulière au cours de ce travail (voir Chap. 6 page 95 et Mahé *et al.*, accepté).

La distribution complexe des lupins de l'Ancien Monde peut vraisemblablement s'expliquer par les facteurs climatiques et humains. Les fluctuations climatiques importantes qui ont marqué les deux derniers millions d'années — pléistocène et holocène (Gradstein & Ogg, 2004) — ont fait se succéder les épisodes secs et humides en Afrique (Douady *et al.*, 2003), et les glaciations au Nord de la Méditerranée (d'Errico & Sánchez Goñi, 2003). Pour les populations de lupins, cette alternance de phases d'extension et de replis dans des zones-refuges (Schmitt, 2007) a pu engendrer la distribution très fragmentée constatée aujourd'hui. À cela s'ajoute la présence humaine — depuis près de 200 000 ans dans cette région si l'on se limite à *Homo sapiens sapiens* (McDougall *et al.*, 2005) —, avec les conséquences qu'elle a pu avoir sur le milieu : anthropisation, éradications et introductions d'espèces.

3.3 Apports récents sur la phylogénie des lupins

Bien qu'ayant reçu l'attention de la communauté scientifique, la phylogénie du genre *Lupinus* pose encore un certain nombre de difficultés. Différentes approches phénétiques ont été testées pour clarifier la taxonomie, l'histoire et les relations entre les lupins : morphologie (Gladstones, 1974, 1998), micromorphologie des graines (Heyn & Herrnstadt, 1977 ; Plitmann & Heyn, 1984 ; Aïnouche & Bayer, 2000 ; Aïnouche *et al.*, 2004 ; Piotrowicz-Cieślak *et al.*, 2008), cytogénétique et croisements (Plitmann & Pazy, 1984 ; Kazimierski, 1988 ; Carstairs *et al.*, 1992 ; Gupta *et al.*, 1996 ; Naganowska *et al.*, 2003), diversité des protéines de stockage et sérologie (Cristofolini, 1989 ; Salmanowicz & Przybylska, 1994 ; Przybylska & Zimniak-Przybylska, 1995 ; Aïnouche, 1998), étude des isozymes (Wolko & Weeden, 1990a, 1990b), flavonoïdes (C. A. Williams *et al.*, 1983) et alcaloïdes (Nowacki, 1963 ; Wink *et al.*, 1995 ; Aïnouche & Bayer, 1996 ; Aïnouche *et*

11. Nikolaï Vavilov (1887–1943) est un botaniste russe qui constitua entre 1916 et 1932 une très importante collection de plantes. Sur la base de ses observations, il fit l'hypothèse que les principales espèces cultivées sont originaires de neuf grands centres de diversité. L'identification de ces régions est capitale pour l'amélioration génétique des espèces et la création de nouvelles variétés, puisqu'elles constituent des gisements de caractères d'intérêt (résistances aux maladies, rusticité, *etc*). En URSS, les années 30 voient l'ascension politique du sélectionneur de plantes Trofim Lyssenko, aux idées en phase avec le régime stalinien. Opposé aux thèses scientifiques de Lyssenko, Vavilov est arrêté par la police politique en 1940. Il est condamné à vingt ans de travaux forcés et meurt de malnutrition en janvier 1943.

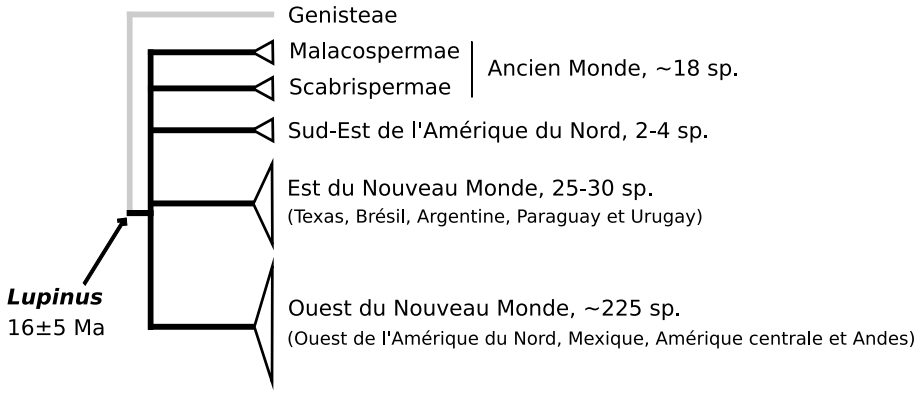

FIG. 3.8 — Principales relations phylogénétiques au sein du genre *Lupinus*, obtenues par l'analyse des séquences ITS (Käss & Wink, 1997a ; Aïnouche *et al.*, 2004) et LEGCYC1A (Hughes & Eastwood, 2006), repris de Stępkowski *et al.* (2007). Le regroupement des espèces de l'Ancien Monde en un clade unique est suggéré par les données mais n'est pas soutenu par les tests statistiques. C'est également le cas pour la phylogénie publiée par Eastwood *et al.* (2008), qui bien que basée sur la combinaison de gènes nucléaires — ITS, LEGCYC1A, LEGCYC1B et GPAT1 — et chloroplastiques — *trn*S-*trn*G, *trn*T-*trn*L, intron du *trn*L et *trn*L-*trn*F —, ne parvient pas à démêler les relations entre les grands groupes phylogénétiques. L'âge du clade *Lupinus* a été estimé par Hughes & Eastwood (2006) sur la base de données fossiles tirées de Lavin *et al.* (2005).

al., 2004). Bien qu'ayant apporté d'importantes connaissances, notamment sur les lupins de l'Ancien Monde, ces approches ne sont pas parvenues à complètement clarifier la situation.

Depuis les années 1990, l'utilisation de séquences nucléotidiques (Käss & Wink, 1997a ; Aïnouche *et al.*, 2004 ; Hughes & Eastwood, 2006 ; Ch. S. Drummond, 2008 ; Eastwood *et al.*, 2008) a fait progresser la classification des lupins et a permis d'évaluer l'âge du clade à $16,01 \pm 5,15$ millions d'années (Hughes & Eastwood, 2006), sans apporter toutefois de réponse définitive (voir Fig. 3.8). Les phylogénies du genre *Lupinus* manquent encore de résolution pour les nœuds les plus profonds de l'arbre, c'est-à-dire les relations entre les grands groupes géographiques.

La très grande diversité des lupins de l'Ouest du Nouveau Monde, est liée à l'histoire géologique et à l'évolution des massifs montagneux (cordillère des Andes, Sierra Madre, Sierra Nevada, chaînes du Pacifique, Montagnes Rocheuses, chaîne des Cascades et chaîne d'Alaska). Le cas des lupins andins a fait l'objet d'études particulières. La diversification de ces lupins est liée au processus géologique de surrection des Andes (Haffer, 2008, pour une revue de l'histoire du bassin amazonien). Le dernier ancêtre commun des 81 espèces actuelles ayant été estimé à moins de 2 millions d'années, il s'agirait d'une radiation très rapide (Hughes & Eastwood, 2006 ; Eastwood & Hughes, 2008)[12].

12. En employant des méthodes statistiques plus évoluées, Moore & Donoghue (2009) ne trouvent pas de corrélation forte entre la dispersion des lupins andins et un rythme de diversification accru. À noter également, Egan & Crandall (2008) signalent chez les fabacées du Nouveau Monde un autre exemple de radiation liée cette fois à un assèchement du climat.

3.4 Des multiples intérêts du lupin

Une plante pionnière Les populations sauvages de lupins colonisent un large spectre d'habitats : du climat sub-arctique de l'Alaska à la forêt tropicale sud-américaine, en passant par le climat méditerranéen et celui des hauts-plateaux andins. Le système racinaire des lupins est large et profond, adapté à des sols pauvres en nutriments et en eau. Cette capacité à exploiter les sols inhospitaliers découle, comme pour les autres légumineuses, d'une symbiose avec des rhizobiums, un groupe de bactéries fixatrices d'azote, appartenant dans le cas des lupins au groupe des *Bradyrhizobium* sp. Ces qualités font des lupins des plantes pionnières des sols mis à nu[13]. C'est par exemple le cas de *L. lepidus* qui a colonisé rapidement les flancs du mont Saint Helens (chaîne des Cascades, État de Washington, États-Unis) après sa dernière éruption majeure en mai 1980 (Findley, 1981 ; Morris & Wood, 1989 ; Bishop & Schemske, 1998).

Une source de protéine La symbiose avec *Bradyrhizobium* permet un métabolisme azoté important dont les excès sont stockés sous forme de protéines. L'archéologie nous montre que différentes espèces de lupins ont très tôt fait partie de l'alimentation humaine. Des traces de consommation ont été retrouvées en ancienne Égypte, et en Amérique chez les différentes civilisations d'Amérique centrale et de la cordillère des Andes (*L. mutabilis*, la seule espèce du Nouveau Monde a avoir été domestiquée) (Kasprzak *et al.*, 2006 ; Eastwood & Hughes, 2008). Le pourtour méditerranéen n'est pas en reste, puisque *L. albus*, originaire des Balkans, est cultivé dans la région méditerranéenne depuis plus de 4 000 ans (Gladstones, 1974, 1984 ; Gross, 1986 ; Cowling *et al.*, 1998). Pour les premiers agriculteurs, la capacité des légumineuses à enrichir les sols en azote a permis la culture de plantes plus exigeantes. Cette stratégie d'alternance des cultures[14] est déjà présente dans les textes anciens (Virgile, I^er^ siècle av. J.-C.).

Vers la sélection de variétés douces En 1781, Frédéric II de Prusse (1740–1786), à la recherche de plantes candidates pour la rotation des cultures, tente d'acclimater le lupin blanc aux sols du Nord de l'Allemagne. Malgré une décennie de tests, le mûrissement tardif des graines de *L. albus* fait échouer la tentative. Plus tard, dans les années 1860, *L. luteus* et *L. angustifolius* sont introduits avec succès et colonisent

13. La résistance des graines de lupins a également été mise en avant. En 1954, des graines de *Lupinus arcticus* sont découvertes lors de fouilles archéologiques dans des terriers de rongeurs du Yukon. Sur la base de leur environnement direct — par stratigraphie notamment —, l'âge de ces graines a été estimé à 10 000 ans. Certaines de ces graines ayant pu germer et se développer normalement (Porsild *et al.*, 1967), elles ont été considérées comme un exemple remarquable de la capacité de conservation du permafrost (Kjøller & Ødum, 1971), jusqu'à ce qu'une radio-datation au carbone-14 révèle qu'il s'agissait de graines modernes ayant contaminé le site archéologique (Zazula *et al.*, 2009).

14. Au cours du XX^e^ siècle, l'arrivée d'un pétrole bon marché a permis l'émergence de l'industrie pétro-chimique et la fabrication massive d'engrais azotés, éliminant le besoin immédiat de rotation des cultures, favorisant la mise en place de grandes mono-cultures céréalières et diminuant l'attrait pour la culture de légumineuses. Ce mode d'agriculture, principalement dédié à l'alimentation animale — et plus récemment à la synthèse d'agrocarburants —, a des conséquences importantes sur la structure et la fertilité des sols, sur la biodiversité et sur la résilience des milieux cultivés.

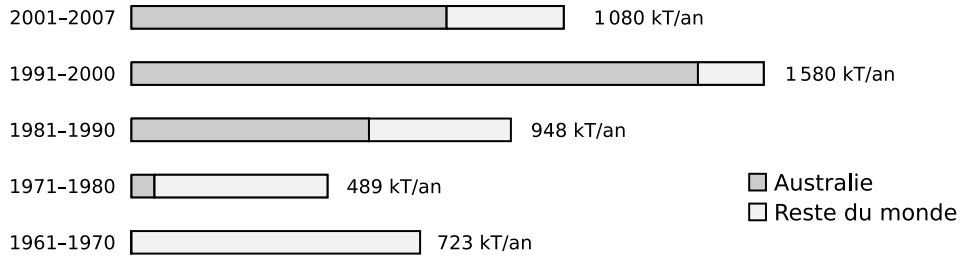

FIG. 3.9 — Évolution de la production mondiale de lupins, en milliers de tonnes par an (kT/an) et moyennée sur dix ans, d'après les données de l'Organisation des Nations unies pour l'alimentation et l'agriculture. L'Australie est depuis le début des années 1980 le premier pays producteur de lupins, malgré une crise sévère due à l'arrivée de l'anthracnose en 1996.

une grande partie des sols sablonneux et acides des plaines côtières de la Baltique. Ils sont alors utilisés comme fourrage et comme fertilisant. Les premières tentatives de passage de formes sauvages ou semi-domestiques à des formes cultivables à grande échelle datent du début du XX^{e} siècle. C'est la première guerre mondiale qui pousse les agronomes allemands à rechercher une plante à graines riches en protéines. Ils se tournent vers *L. albus* et *L. luteus* et cherchent à en réduire la teneur en alcaloïdes, désagréables au goût et neurotoxiques à forte dose[15]. Dans les années qui suivent, des chercheurs australiens entreprennent la domestication de *L. angustifolius* et obtiennent en 1967 une espèce cultivable à grande échelle, adaptée aux conditions climatiques et édaphiques de l'Ouest australien[16], et à faible teneur en alcaloïdes : initialement 0,01 à 0,03 % puis 0,001 à 0,006 % aujourd'hui, contre 0,8 à 0,9 % pour les variétés sauvages (Gill & Vear, 1980 ; Clements *et al.*, 2008).

Passage à une échelle industrielle Entre 1960 et 1980, la production mondiale de lupin est dominée par le bloc soviétique, notamment la Pologne et l'ex-Allemagne de l'Est, avec une production de 600 000 tonnes par an en moyenne. Dans les années 1980 et 1990, l'Australie devient le premier producteur mondial et porte la production annuelle à près de 2 millions de tonnes (voir Fig. 3.9)[17].

Aujourd'hui l'essentiel de la production australienne est composé de *Lupinus angustifolius* et de *L. albus*. La production a atteint un pic en 1996 coïncidant avec l'arrivée sur le sol australien de l'anthracnose (*Colletotrichum lupini*). Des variétés résistantes ont été rapidement créées pour *L. angustifolius* et *L. albus* mais pas pour *L. luteus*, cantonnant cette espèce aux régions dans lesquelles *L. cosentinii*, l'hôte naturel de l'anthracnose, n'est pas présent (Adhikari *et al.*, 2008 ; Shea *et al.*, 2008). Malgré des mesures de contrôle drastiques, les pathogènes — rouille du lupin (*Uromyces lupinicola*), *Fusa-*

15. La teneur maximale en alcaloïdes est fixée à 200 mg/kg d'aliment (Resta *et al.*, 2008).

16. Il est à noter qu'en Australie, les bactéries nécessaires à la nodulation ont été cultivées par l'industrie et inoculées dans les sols destinés à recevoir les lupins. Cette pratique n'est plus nécessaire puisque la souche de bactérie introduite accidentellement avec les premiers lupins a colonisé la totalité du Sud-Ouest australien (Howieson & O'Hara, 2008).

17. `http://faostat.fao.org/site/408/DesktopDefault.aspx?PageID=408` (consulté en mai 2009).

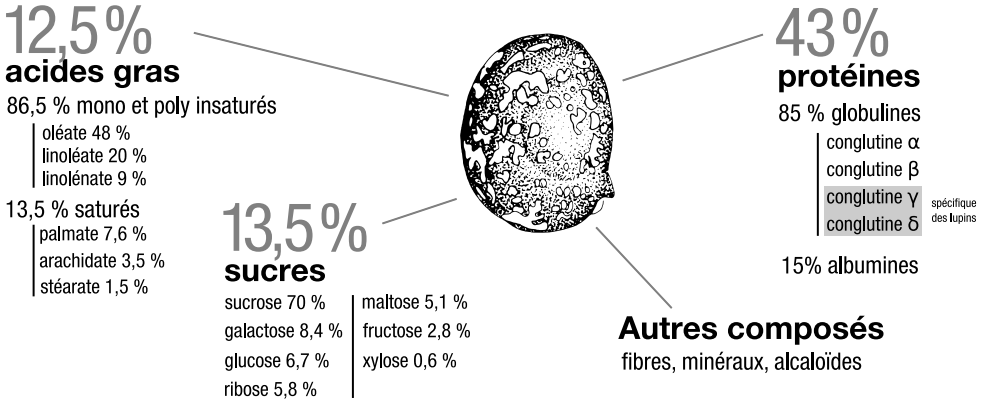

FIG. 3.10 — Composition moyenne d'une graine de lupin. L'enveloppe de la graine, riche en fibres, représente environ 30 % du poids total (Sipsas, 2008). Les protéines de réserve sont en majorité des conglulines, les formes γ et δ étant spécifiques des lupins. La graine contient également des minéraux (5,5 %), des alcaloïdes — lupanine, spartéine, angustifoline, cytosine —, des phytates (composés phosphorés), des tanins et des oligosaccharides (Petterson, 1998). Les valeurs sont données à titre indicatif, les teneurs pouvant varier de façon importante entre deux lignées.

rium, anthracnose, *Diaporthe toxica* — qui sévissent en Europe, atteignent l'Australie et mettent en danger l'agriculture locale (McKirdy *et al.*, 2008). La Russie, autrefois grand producteur de lupins, a elle aussi été touchée massivement par l'anthracnose entre 2003 et 2006 (Yakusheva & Svist, 2008).

En Europe, jusque récemment, la majorité de la production était composée de *L. luteus* ou *L. albus*. Mais la propagation du *Fusarium* a considérablement réduit la production européenne, le temps que des variétés résistantes soient créées (Buirchell, 2008). Ces crises ont là aussi favorisé l'espèce *L. angustifolius*. Au cours des dix dernières années, la production française a augmenté, avec une dominance de *L. albus*. On retrouve cette préférence pour *L. albus* (mais aussi *L. luteus*) dans la production de plusieurs pays du pourtour méditerranéen (Maroc, Portugal, Espagne, Italie, Grèce et Égypte)[18].

En Amérique du Sud, c'est le Chili qui est devenu en quelques années le principal producteur local, avec l'espèce *L. albus* originaire de l'Ancien Monde. Mais les cultures de lupins y sont aujourd'hui en compétition avec les cultures céréalières dédiées à la production d'agrocarburants, plus rentables à court terme (Baer, 2008). De petites quantités de *L. mutabilis* sont traditionnellement cultivées par les indiens natifs d'Équateur, du Pérou et de Bolivie.

Intérêts nutritionnels Les graines de lupins ont une forte teneur en protéines (43 %) et contiennent également des fibres (25,5 %), des sucres (13,5 %), des matières grasses (12,5 %) et des minéraux (5,5 %) (voir Fig. 3.10). Elles constituent donc un aliment de choix pour l'alimentation des animaux d'élevage (Berk *et al.*, 2008 ; Fychan *et al.*, 2008 ; Marley *et al.*, 2008 ; Serrano *et al.*, 2008 ; Yáñez-Ruiz *et al.*, 2009) et un bon complé-

18. Ces pays produisent des variétés plus rustiques de *L. albus*, plus chargées en alcaloïdes.

ment pour l'alimentation humaine (Edwards & Barneveld, 1998). Selon U. Hennig *et al.* (2008), une prise alimentaire composée de 17 % de graines de lupins, 62 % de blé et 17 % de poisson séché couvre les recommandations de l'Organisation mondiale pour la santé, y compris pour les huit acides aminés essentiels[19]. L'huile contenue dans les graines de lupins est d'excellente qualité nutritionnelle, avec une forte proportion d'acides gras insaturés et un ratio oméga-3/oméga-6 très favorable (Boschin *et al.*, 2008 ; Suchý *et al.*, 2008). Les graines de lupins sont également riches en lutéine (caroténoïdes), un composé permettant de lutter contre l'apparition de la dégénérescence maculaire liée à l'âge[20] (Fryirs *et al.*, 2008). La consommation de lupin, comme celle d'autres légumineuses, augmente la sensation de satiété (Hodgson & Lee, 2008), a un effet hypoglycémique et hypocholestérolémique (Garzón-de la Mora, Moreno-Sandoval *et al.*, 2008 ; Garzón-de la Mora, Villáfan-Bernal *et al.*, 2008 ; Gurrola-Díaz *et al.*, 2008) et participe à la diminution du risque d'accidents cardio-vasculaires (C. R. Sirtori *et al.*, 2004 ; Martins *et al.*, 2005 ; Y. P. Lee *et al.*, 2009), avec l'avantage de présenter peu de composants anti-nutritionnels (E. Sirtori *et al.*, 2008) et peu de phytoœstrogènes (Arnoldi, 2008).

Apparition d'allergie En raison du renchérissement des produits céréaliers et de la demande en protéines végétales « non-OGM », l'industrie alimentaire s'est tournée vers le lupin. Avec cette utilisation accrue de farine de lupin, notamment comme épaississant et émulsifiant (Dijkink *et al.*, 2008), des cas d'allergie sont apparus (Fæstea *et al.*, 2004 ; Brennecke *et al.*, 2007 ; Quaresma *et al.*, 2007 ; Wassenberg & Hofer, 2007 ; Reis *et al.*, 2007 ; Peeters *et al.*, 2007) et la Commission européenne a inscrit en 2006 le lupin sur la liste des allergènes connus, obligeant les producteurs à signaler sa présence sur l'étiquette de tout produit alimentaire[21]. La caractérisation des protéines allergènes à commencé (P. M. C. Smith *et al.*, 2008 ; Kłos *et al.*, 2008) et des méthodes de détection de traces d'ADN de lupin dans la nourriture sont en développement (Demmel *et al.*, 2008).

Autres pistes biotechnologiques Les lupins présentent également d'intéressantes capacités de détoxication des sols contaminés par des métaux lourds (cadmium, cobalt, manganèse, nickel, plomb, zinc) (Petterson & Harris, 1995 ; Page *et al.*, 2006 ; Peñalosa *et al.*, 2007), par des hydrocarbures (Dashti *et al.*, 2009) ou par des produits chimiques (Bonvallot, 2004). Les recherches s'orientent également vers une sélection de

19. Le métabolisme humain nécessite huit acides aminés essentiels : le tryptophane, la lysine, la méthionine, la phénylalanine, la thréonine, la valine, la leucine et l'isoleucine. Deux autres, l'histidine et l'arginine sont dits semi-essentiels car seuls les nourrissons ont besoin d'un apport exogène.

20. La dégénérescence maculaire liée à l'âge est une maladie incurable touchant 30 % des européens de plus de 75 ans. La lutéine protège la macula en filtrant les longueurs d'ondes courtes, à forte énergie, responsables de la formation de radicaux libres.

21. Commission Directive 2006/142/EC of 22 December 2006 amending Annex IIIa of Directive 2000/13/EC of the European Parliament and of the Council listing the ingredients which must under all circumstances appear on the labelling of foodstuffs. (2006). *Official Journal — European Union Legislation, 49(368)*, 110-111.

bactéries directement capables de dégrader ou de neutraliser les contaminants. Les lupins et les autres légumineuses — naturellement associées à des bactéries — sont donc les candidats idéaux pour accueillir ces bactéries modifiées (Barac *et al.*, 2004).

Enfin, la large gamme d'alcaloïdes, de phytoalexines et de flavonoïdes produite par les lupins pourrait être utilisée en médecine ou comme produit phytosanitaire (Folkman *et al.*, 2002). En effet, les alcaloïdes de lupins ont un effet répulsif sur les mollusques (Wink, 1984c), les bactéries (Wink, 1984b), les champignons (Zamora-Natera *et al.*, 2008) et les insectes (Wink, 1984a, 1992).

Deuxième partie

Méthodologie

4

De la mise en culture au séquençage

On décrira ici le matériel végétal et les différentes méthodes utilisées, depuis la mise en culture, l'analyse de la microstructure des téguments séminaux, la cytogénétique moléculaire, le séquençage de l'ADN, l'évaluation des séquences et leur traitement en vue de leur utilisation dans les analyses phylogénétiques et génomiques.

4.1 Matériel végétal et mise en culture

Notre échantillonnage couvre la quasi-totalité des lupins de l'Ancien Monde et inclut 48 taxons représentatifs des différents groupes géographiques du Nouveau Monde. Il inclut également l'espèce *Lupinus mariae-josephi* H. Pascual (2004) nouvellement décrite, ainsi que plusieurs autres espèces, sous-espèces ou populations au statut taxonomique incertain. Onze taxons externes (*outgroups*) ont également été utilisés (voir Tab. D page 187). Les échantillons sont issus de collectes sur le terrain, de collections vivantes disponibles au laboratoire ou d'herbiers[1]. Les échantillons ont été déterminés sur la base d'ouvrages de référence, dont Gladstones (1974, 1984, 1998), Aïnouche (1998, et références citées) et *Flora europaea* (Tutin *et al.*, 1968-1980).

Analyse des microstructures de graines Les graines sont nettoyées à l'éthanol 95 %, séchées puis montées sur un support en aluminium à l'aide de ruban adhésif double-face. Pour permettre la transmission des électrons, les échantillons sont métallisés par

1. À noter que certains échantillons sont issus d'herbiers anciens, dont celui d'Augustin Pyrame de Candole (1778–1841). Né à Genève, il étudie puis enseigne en France. En 1806, il reçoit la mission de parcourir tout l'Empire pour reconnaître l'état de l'agriculture. En 1813, il fait paraître la *Théorie élémentaire de la botanique*. La Restauration l'oblige à se réfugier à Genève en 1816. Il y entreprend en 1818 une description de toutes les plantes connues, et publie les deux premières parties de ce travail (*Regni vegetabilis systema naturale*, 1818–1821). Il poursuit dans un ouvrage plus abrégé, *Prodromus regni vegetabilis*, continué après sa mort par son fils Alphonse Louis Pierre Pyrame de Candolle (1806–1893) (14 vol. in-8, 1824–1862). Cet ouvrage immense décrit 90 000 plantes.

pulvérisation cathodique (Jeol JFC 1100) avant d'être observés au microscope électronique à balayage Jeol JSM-6301[2]. Pour chaque échantillon, le tégument est examiné près du centre de la graine, près du hile et en coupe transversale. Les observations ont été comparées à celles de la littérature (Heyn & Herrnstadt, 1977 ; Bragg, 1983 ; Plitmann & Pazy, 1984 ; Monteiro, 1987 ; Aïnouche, 1998 ; Aïnouche & Bayer, 2000 ; Aïnouche *et al.*, 2004).

Germination Les graines de lupins sont mises à germer sur papier filtre imbibé d'eau distillée. Il a été montré que les composés relâchés par les graines peuvent inhiber le développement des bactéries *Bradyrhizobium* et de fait diminuer les chances que l'association symbiotique se fasse correctement (Abd-Alla, 1998). Quelques heures dans l'eau suffisent pour nettoyer les graines et éliminer la plus grande partie des composés hydrosolubles. Les plantules sont ensuite transférées dans un petit volume de terreau ou de support neutre. Après développement des premières feuilles, les lupins sont cultivés sous serre, dans des pots contenant un mélange composé d'1/4 de sable, d'1/4 de terre et d'1/2 de terreau. Selon les besoins de l'étude, des pointes de racines sont prélevées sur des germinations de 24 à 36 heures pour la cytogénétique, et des feuilles sont prélevées sur des plantules ou des jeunes plants de 2 à 4 semaines pour la cytométrie en flux et les extractions d'ADN.

4.2 Cytogénétique moléculaire

Au cours de la division cellulaire, l'ADN se condense en chromosomes et l'enveloppe nucléaire disparaît. Les chromosomes condensés se rassemblent ensuite à l'équateur de la cellule pour former la plaque équatoriale (métaphase). C'est à ce stade qu'ils sont le plus facilement observables et que l'on peut les compter. Les pointes racinaires étant en phase de croissance, ce sont d'excellentes candidates pour observer des cellules en cours de division. Les méthodes détaillées ci-dessous ont été adaptées des protocoles n^{os} 5, 9, 13, 14 et 22 de l'UMR « Amélioration des plantes et biotechnologies végétales » de l'INRA du Rheu. Pour une approche plus générale de l'hybridation *in situ* (*Fluorescent in situ hybridization* ou FISH), on consultera Schwarzacher & Heslop-Harrison (2000).

Préparation des racines Les graines sont scarifiées puis mises en germination sur papier humide pendant 48-72 heures. Les extrémités de racines sont sectionnées (3-5 mm) et placées 4 heures dans une solution aqueuse de 8-hydroxyquinolénine à 0,04 % (290 mg/l) pour bloquer les cellules en métaphase I. Les divisions cellulaires sont ensuite fixées par un séjour de 48 heures dans une solution de Farmer[3]. Les pointes de racines sont conservées à -20 °C dans de l'éthanol 70 %.

2. Centre de microscopie électronique à balayage et micro-analyse, université de Rennes 1.
3. Trois volumes d'éthanol pour un volume d'acide acétique glacial (3:1).

Préparation des lames Les pointes de racines sont rincées 2 × 10 minutes dans de l'eau ultra-pure afin d'éliminer le fixateur, puis 15 minutes dans du tampon citrate 0,01 M pH 4,5[4]. Les racines sont ensuite digérées 45-60 minutes dans 250 µl de solution enzymatique[5] à 37 °C. Les racines sont rincées au moins 30 minutes dans de l'eau ultra-pure puis transférées sur une lame. L'excès d'eau est éliminé à l'aide d'un papier absorbant puis la racine est hachée dans une goutte d'éthanol-acétique (3:1) et étalée sur la lame. Après séchage à l'air libre, la lame est colorée par une goutte de DAPI[6] (4',6' di amidino-2-phényl indole). La lame est conservée à l'obscurité à −20 °C ou observée immédiatement à un grossissement de 1 000× sous un microscope Olympus BX-51 à épifluorescence muni d'une caméra numérique Pixera Penguin 600ES refroidie à −20 °C et couplée au logiciel IMAGE-PRO 5.0. À ce stade, l'observation au microscope permet de sélectionner les lames les plus riches en « plaques métaphasiques » et de réaliser un comptage chromosomique. Les lames sélectionnées peuvent ensuite être hybridées avec une ou plusieurs sondes afin de localiser *in situ* les portions de génome correspondantes (FISH). Notre objectif étant d'évaluer le degré d'envahissement des génomes par les rétrotransposons, nous avons utilisé — en plus de la sonde-témoin ARN 45S — des sondes caractéristiques de ces éléments.

Préparation de la sonde Les sondes utilisées sont des séquences de transcriptase inverse (*reverse transcriptase* ou *rt*) de rétrotransposons Ty1 et Ty3 (voir section 4.5.5 page 67). Dans un premier temps, nous avons utilisé des sondes marquées à la biotine (vitamine H) selon la méthode d'« amorçage aléatoire » (*random priming*) décrite par Salvo-Garrido *et al.* (2001), et à l'aide du kit *Bioprime DNA Labeling System* d'Invitrogen Fisher Bioblock. Le marquage aléatoire de la sonde ne s'étant pas révélé efficace, nous avons utilisé une méthode basée sur la synthèse de sonde à partir de nucléotides marqués à la digoxigénine-11-2'-déoxy-uridine-5'-triphosphate *PCR DIG Labeling Mix* (Roche). Le marquage aléatoire ajoute une terminaison fluorescente à la sonde tandis que dans le cas du marquage à la digoxigénine, c'est la sonde elle même qui est composée d'éléments fluorescents. La sensibilité est donc bien plus forte. Le protocole d'amplification est identique à celui utilisé pour les *rt* de Ty1 et Ty3 (voir page 67).

Hybridation *in situ* Les lames préparées en amont sont rincées dans du 2× SSC[7] et couvertes par 150 µl d'une solution de RNAse à 0,1 mg/ml de 2× SSC. Le tout est recouvert d'une lamelle plastique et placé dans un incubateur à 37 °C, sous atmosphère humide, pendant une heure. Les lames sont ensuite lavées 3 minutes dans un bain de 2× SSC à 42 °C, sous agitation (étape répétée deux fois). Les lames sont drainées avant d'être traitées à la pepsine pendant 20 minutes à 37 °C, afin d'éliminer les protéines cytoplasmiques résiduelles (pepsine à 100 µg/ml de HCl à 0,01 M, 100 µl/lame). Les lames sont rincées 3 minutes dans un bain de 2× SSC à 42 °C, sous agitation (étape

4. 1,47 g de trisodium citrate-dihydrate et 1,05 g de citrate monohydrate dans 500 ml d'eau ultra-pure.
5. 0,25 g d'Onozuka R-10 cellulase et 0,05 g d'Y23 pectolyase dans 5 ml de tampon citrate.
6. Molécule fluorescente (couleur bleue) capable de se lier fortement à l'ADN.
7. Le 20× SSC est composé de 175,3 g de NaCl et de 88,2 g de trisodium citrate dissous dans un volume final d'un litre d'eau ultrapure. Dans ce protocole, il est utilisé en dilution à 4×, 2× et 0,1×.

répétée deux fois). Pour restaurer la structure protéique des chromosomes, les lames sont traitées au paraformaldéhyde (0,05 g/ml d'eau + NaOH à 0,01 M) pendant 10 minutes, suivies d'un rinçage de 3 minutes dans un bain de 2× SSC (répété deux fois). L'ADN est dénaturé dans un bain de formamide à 70 % (déionisé par 5 g de résine échangeuse d'ions) pendant 2 minutes à 70 °C. Les lames sont déshydratées par passages de 3 minutes dans des solutions d'alcool à 70 %, 90 % et 100 % à −20 °C, puis séchées à température ambiante pendant 2 heures.

Le solution d'hybridation est la suivante : 25 µl de formamide 100 %, 10 µl de sulfate de dextran à 0,5 g/ml, 5 µl de 20× SSC, 0,6 µl de dodécylsulfate de sodium (détergent) à 0,2 g/ml, 5 µl d'Alexa vert (ARNr 45S), 100 ng de sonde, 10 µg d'ADN de saumon soniqué et le volume d'eau ultrapure nécessaire pour atteindre un volume final de 50 µl. Une sonde correspondant à l'ARN 45S (unité de transcription regroupant le 18S, l'ITS1, le 5,8S, l'ITS2 et le 26S) a été utilisée comme témoin positif, pour identifier les *loci* d'ARNr et vérifier le bon déroulement de l'étape d'hybridation. La solution est dénaturée par un séjour de 6 minutes à 92 °C, puis immédiatement transférée sur de la glace (15 minutes). Les lames sont recouvertes de 50 µl de la solution d'hybridation et d'une lamelle en plastique, avant d'être placées dans un incubateur à 37 °C, sous atmosphère humide, pendant environ 12 heures.

Pour détecter le marquage à la digoxygénine, 40 µl de solution d'anti-Dig-FITC à 200 µg/ml (Roche) sont ajoutés à 340 µl de *Bovine serum albumin* à 5 %. Les lames sont recouvertes de 50 µl de la solution de détection et d'une lamelle en plastique, avant d'être placées dans un incubateur à 37 °C, sous atmosphère humide, pendant une heure. Les lames sont ensuite rincées 5 minutes dans un bain de 4× SSC-Tween[8] à 42 °C, sous agitation (étape répétée trois fois). Les lames sont drainées puis colorées par une goutte de DAPI avant d'être observées à trois longueurs d'ondes différentes (filtres à 461, 488 et 568 nm). Ces expériences de cytogénétique moléculaire ont été réalisées dans l'unité mixte de recherche « amélioration des plantes et biotechnologies végétales » de l'INRA du Rheu, en collaboration avec Olivier Coriton et Virginie Huteau.

4.3 Cytométrie en flux

Cette technique consiste à doser la quantité d'ADN directement dans les cellules isolées à partir d'un échantillon de tissu frais selon la méthode de Galbraith *et al.* (1983). Deux standards ont été utilisés : *Lycopersicon esculentum* (2C = 2 pg) et *Petunia hybrida* (2C = 2,85 pg ; Marie & Brown, 1993). Un fragment de feuille du standard et de l'échantillon sont finement hachés ensemble dans 600 µl de tampon osmotique de Galbraith[9] afin d'extraire les noyaux des cellules. La solution est ensuite filtrée, et 30 µl d'iodure de propidium (marqueur fluorescent de l'ADN) à 60 µg/ml et 1 µl de RNAse à 10 mg/ml sont ajoutés au filtrat. Après 10 minutes, le dosage de l'ADN des noyaux est réalisé à l'aide d'un Laser Argon 20 mW à 488 nm. Les dosages ont été

8. 998 ml de 4× SSC + 2 ml de Tween 20 (Merck Eurolab).
9. 1 % de Triton (détergent) + 1 % de polyvinylpyrrolidone (inhibiteur de la polyphénol oxydase).

réalisés au service de cytométrie en flux de l'Institut des Sciences du Végétal, CNRS de Gif-sur-Yvette, en collaboration avec Spencer Brown et Olivier Catrice.

4.4 Extraction d'ADN

Le broyage des tissus végétaux (frais ou déshydratés) est réalisé par la technique du *bead-beating*. L'utilisation d'un broyeur à billes (modèle Retsch MM 301) présente plusieurs avantages par rapport à la technique classique du « mortier-pilon » : pas de contact direct entre l'échantillon et l'azote liquide, temps d'exécution court, pas de perte de matière et surtout reproductibilité expérimentale. Environ 100 mg de matière fraîche ou 20 mg de matière sèche sont placés dans des tubes de 2 ml. Deux billes d'acier de 3 mm de diamètre sont ajoutées dans chaque tube. Après un passage dans l'azote liquide (-180 °C), le broyage se fait en deux séquences de 30 s à une fréquence de 30 cycles/s. Entre les deux séquences et au terme du broyage, les échantillons sont replongés dans l'azote liquide.

L'extraction d'ADN est ensuite réalisée *via* les kits Macherey-Nagel (*NucleoSpin Plant* I et II) ou Qiagen (*kit Plant Tissue* ou *Plant Mini kit*) et permet d'obtenir environ 10 à 50 µg d'ADN. Pour certaines graines trop anciennes pour germer, l'ADN a été extrait par *Kit Food* (Macherey-Nagel), un traitement chimique spécialement dédié aux échantillons très riches en protéines ou en matière grasse. L'ADN obtenu est dégradé et en quantité plus faible, mais il reste utilisable pour des amplifications de séquences de quelques centaines de nucléotides[10]. Pour tous les extraits, la qualité de l'ADN a été contrôlée par migration sur gel d'agarose à 12 g/l (1,2 %) dans un bain de TBE $0,5\times$ (Tris, Borate, EDTA), coloration dans un bain de bromure d'éthidium à 0,5 g/l (agent intercalant ; Sharp *et al.*, 1973), et révélation sous lumière ultraviolette. L'ADN a été quantifié par spectrophotométrie (*Nanodrop*), puis les extraits ont été conservés à -20 °C.

4.5 Gènes étudiés, amorces et amplification

L'émergence de la biologie moléculaire a permis d'accéder à d'immenses gisements de données utilisables pour la reconstruction phylogénétique (Zuckerkandl & Pauling, 1965). Cependant, en fonction de la question biologique posée — phylogénie au sein d'un seul genre ou phylogénie des grands groupes d'angiospermes —, les gènes doivent être choisis pour leur vitesse d'évolution. Historiquement, les gènes chloroplastiques, à évolution lente, ont été très utilisés. L'arrivée d'amorces universelles pour les espaceurs internes et externes de l'ARN ribosomique nucléaire a popularisé leur utilisation (White *et al.*, 1990) et a permis de clarifier la phylogénie dans de nombreux groupes d'angiospermes (Baldwin *et al.*, 1995 ; Linder *et al.*, 2000). Aujourd'hui, notamment pour éviter certains artéfacts liés à la dynamique évolutive des

10. L'extraction et l'utilisation d'ADN ancien pose beaucoup de difficultés, dont la contamination par des champignons, l'hydrolyse des brins d'ADN et les modifications chimiques des bases (Orlando, 2005).

ARN ribosomiques (Álvarez & Wendel, 2003 ; Soltis *et al.*, 2008), les gènes nucléaires reçoivent une attention croissante.

Afin d'affiner le cadre phylogénétique dans lequel les analyses ultérieures vont se placer — variation de taille de génome, diversité des rétrotransposons —, nous avons essayé de prolonger les travaux publiés par Käss & Wink (1997a) ; Aïnouche & Bayer (1999) ; Aïnouche *et al.* (2004) et Hughes & Eastwood (2006) en complétant les jeux de données pour les espaceurs internes transcrits (ITS) et les espaceurs externes transcrits (ETS) de l'ARN ribosomique nucléaire. Nous avons également obtenu des séquences chloroplastiques (*rbc*L et *trn*L-*trn*F) pour certaines espèces de l'Ancien Monde au statut incertain, comme *L. mariae-josephi*. En complément des ITS et ETS, nous avons identifié et séquencé une partie du gène *SymRK*, un élément clé du processus d'association symbiotique plantes/micro-organismes. Enfin, une première évaluation de la diversité des rétrotransposons présents chez les lupins a été faite par séquençage d'une portion de la transcriptase inverse (*rt*) des rétroéléments Ty1 et Ty3.

Conception d'amorces Mis à part l'utilisation d'amorces disponibles dans la littérature, nous avons été amenés à concevoir nous même les amorces permettant d'amplifier certaines régions ciblées chez les lupins ou les génistoïdées telles que les ETS ou le *SymRK* par exemple. Les amorces doivent être choisies à partir de zones conservées, avoir une taille comprise entre 20 et 27 pb, commencer et finir par des GC de préférence et ne pas être riches en AT. La température de fusion[11] doit être comprise entre 45 et 70 °C, et les couples ne doivent pas former d'hétérodimères. Les logiciels PRIMER3PLUS[12] (Rozen & Skaletsky, 2000), PRIMACLADE[13] (Gadberry *et al.*, 2005) ou le très prometteur UNIPRIME[14] (Bekaert & Teeling, 2008) permettent d'automatiser complètement la conception d'amorces. Des banques de candidats possibles sont également maintenues sur le web (PIP[15] ; L. Yang *et al.*, 2007) et des outils comme GEMPROSPECTOR[16] peuvent générer automatiquement des couples d'amorces amplifiant des régions introniques potentiellement riches en signal phylogénétique. Toutes les amorces utilisées au cours de ce travail sont positionnées sur les schémas des régions amplifiées (présentés ci-après) et leurs séquences sont disponibles page 185. Les amorces ont été synthétisées par la société *Operon* (aujourd'hui MWG).

4.5.1 Les régions ITS et ETS de l'ARN ribosomique nucléaire

De nouvelles séquences ITS et ETS ont été amplifiées, clonées (voir *clonage* p. 68) et séquencées en vue de compléter les bases de données déjà disponibles au laboratoire.

11. Le calcul de la température de fusion se base sur la thermodynamique des appariements et mésappariements de paires de bases, données obtenues empiriquement par Allawi & SantaLucia (1997, 1998a, 1998b, 1998c) et Peyret *et al.* (1999). La salinité joue également un rôle important. Le logiciel MELTING (Le Novère, 2001) donne l'équation utilisée et permet de prendre en compte la totalité des paramètres.
12. `http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi`
13. `http://www.umsl.edu/~biology/Kellogg/primaclade.html`
14. `http://code.google.com/p/uniprime/`
15. `http://ibi.zju.edu.cn/pgl/pip/`
16. `http://cgi-www.daimi.au.dk/cgi-chili/GeMprospector/main`

Dans le cas des ITS, les amorces universelles de White *et al.* (1990) et de Baldwin *et al.* (1995) ont été utilisées avec succès chez les lupins (voir Fig. 4.1). Dans le cas des ETS, deux types d'amorces ont été utilisées dans un premier temps : (1) les amorces Gast1 et 18S-IGS de Chandler *et al.* (2001) pour les espèces de lupins de l'Ancien Monde ; (2) tandis que des amorces anti-sens plus spécifiques, Lu1-Gast1 et Ge1-Gast1, ont été mises au point au laboratoire pour, respectivement, l'amplification des ETS des lupins du Nouveau Monde et celle de certaines génistées, selon le protocole décrit par Baldwin & Markos (1998). Dans un deuxième temps, l'amorce 281F (Cubas & Pardo, communication personnelle) a été utilisée sur 65 de nos échantillons pour amplifier une portion plus grande de la région ETS.

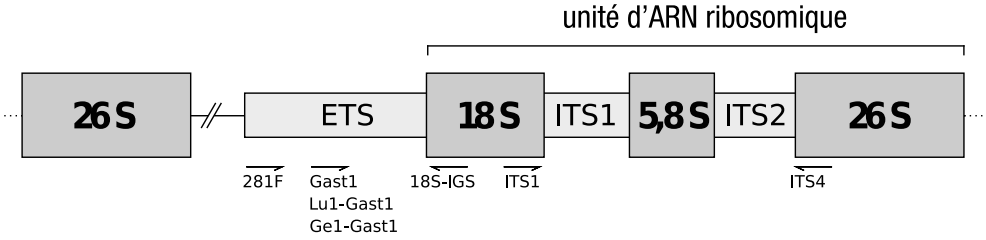

FIG. 4.1 — Schéma d'une répétition en tandem du gène de l'ARN ribosomique nucléaire ; emplacement des différentes régions amplifiées et des amorces correspondantes (l'amorce Gast1 peut être remplacée par Lu1-Gast1 ou Ge1-Gast1, en fonction du groupe visé). Échelle non-respectée.

La solution d'amplification — identique pour les ITS et les ETS — consiste en 5 µl de tampon 10×, 5 µl de dNTPs à 2 mM/chacun, 2 µl de chaque amorce à 5 µM, 2,5 unités de Red Taq polymerase Sigma, 10 à 50 pg d'ADN matrice, dans un volume réactionnel final de 50 µl. Après dénaturation à 94 °C pendant 3 minutes, l'amplification est réalisée en 35 cycles comportant chacun 45 secondes à 94 °C, 45 secondes à 48 °C et 1 minute à 72 °C, suivis d'une élongation finale de 7 minutes à 72 °C. Après vérification sur gel, les produits de PCR sont purifiés *via* le kit BioBasic EZ10 *PCR Purification* et quantifiés par spectrophotométrie.

4.5.2 Les régions chloroplastiques *rbc*L et *trn*L-*trn*F

Le gène *rbc*L Ce gène chloroplastique mono-exonique code pour la grande sous-unité de la ribulose 1,5 bisphosphate carboxylase/oxygénase (RuBisCO, pour revue : Andersson & Backlund, 2008). La séquence de ce gène est faiblement variable et permet de mettre en évidence les relations phylogénétiques entre grands groupes de plantes (Chase *et al.*, 1993). Il n'apporte que peu d'information phylogénétique pour les niveaux infra-génériques. Cependant, l'ADN chloroplastique étant transmis uniquement par la « lignée maternelle », nous l'avons utilisé pour tenter de déterminer la position phylogénétique de *Lupinus mariae-josephi* et de mettre en évidence une éventuelle origine hybride de cette espèce nouvellement décrite (voir Mahé *et al.* ; page 169).

Une portion d'environ 1 400 pb couvrant la quasi-totalité du gène *rbc*L est amplifiée

à l'aide des amorces N et R (Käss & Wink, 1997b). La solution d'amplification consiste en 7 µl de tampon 10×, 2,5 µl de dNTPs à 2 mM/chacun, 5 µl de chaque amorce à 5 µM, 2 unités de Red Taq polymerase Sigma, 10 à 50 pg d'ADN matrice, dans un volume réactionnel final de 50 µl. Après dénaturation à 94 °C pendant 2 minutes, l'amplification est réalisée en 30 cycles comportant chacun 30 secondes à 94 °C, 30 secondes à 45 °C et 1 minute à 72 °C, suivis d'une élongation finale de 10 minutes à 72 °C. Après vérification sur gel, les produits de PCR sont purifiés, quantifiés et clonés.

La région *trn*L-*trn*F Il s'agit de l'intron séparant les deux exons codants pour l'ARN de transfert de la leucine (*trn*L), et de l'espaceur intergénique (IGS) situé entre l'exon 3' du *trn*L et l'unique exon du *trn*F, codant pour l'ARN de transfert de la phénylalanine (F) (voir Fig. 4.2). Comme pour le *rbc*L, nous avons choisi d'amplifier ces deux régions pour essayer de clarifier la situation du taxon *L. mariae-josephi*. L'intron du *trn*L, d'environ 540 pb, est amplifié à l'aide des amorces C et D et l'espace intergénique (385 pb) a été amplifié à l'aide des amorces E et F (Taberlet *et al.*, 1991). La solution d'amplification consiste en 7 µl de tampon 10×, 5 µl de dNTPs à 2 mM/chacun, 5 µl de chaque amorce à 5 µM, 2 unités de Red Taq polymerase Sigma, 10 à 50 pg d'ADN matrice, dans un volume réactionnel final de 50 µl. Après dénaturation à 94 °C pendant 2 minutes, l'amplification est réalisée en 30 cycles comportant chacun 1 minute à 94 °C, 1 minute à 48 °C et 2 minutes à 72 °C, suivis d'une élongation finale de 10 minutes à 72 °C. Après vérification sur gel, les produits de PCR sont purifiés, quantifiés et clonés.

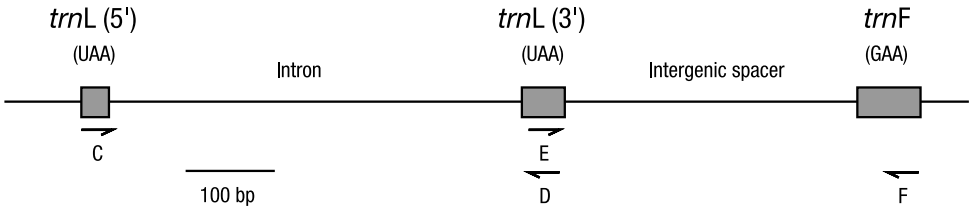

FIG. 4.2 — Structure de la région chloroplastique codant pour les ARN de transfert L et F, et positions des amorces utilisées (d'après Taberlet *et al.*, 1991).

4.5.3 Le gène LEGCYC1A

Chez les angiospermes, les produits des gènes *Cycloidea* sont impliqués dans l'architecture florale. Chez les légumineuses, ces gènes (*Leguminous Cycloidae*, LEGCYC) se divisent en deux classes, la première classe comptant elle-même deux gènes paralogues : LEGCYC1A et LEGCYC1B (Citerne *et al.*, 2003). Ces gènes présentent deux motifs hautement conservés — TCP et R — encadrés par des zones très variables d'un genre à l'autre (Citerne, 2005). Nous nous intéressons en particulier à la copie LEGCYC1A pour lequel nous disposons d'amorces spécifiques dans les régions externes et d'amorces spécifiques dans les régions internes. Selon Ree *et al.* (2004), la duplication du gène LEGCYC1 et la redondance fonctionnelle qu'elle entraîne pourrait expli-

quer la vitesse d'évolution plus rapide de la copie A par rapport à la copie B, comme postulé par Ohno (1970).

Hughes & Eastwood (2006) ont utilisé la partie 5′ du paralogue 1A, incluant le motif TCP, pour essayer de clarifier la phylogénie des lupins américains. Nous avons utilisé les séquences disponibles sur GenBank pour tester la position phylogénétique de *Lupinus mariae-josephi*.

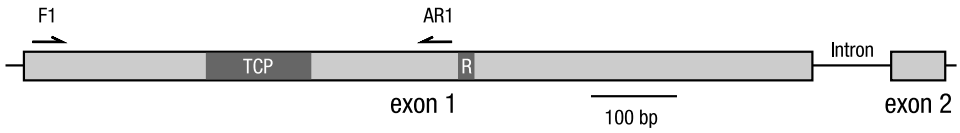

FIG. 4.3 — Structure du gène LEGCYC1A et position des amorces utilisées (d'après Citerne, 2005). Le gène LEGCYC1A est composé de 2 exons, le premier exon comprenant deux régions très conservées : TCP et R. Le domaine TCP est un motif hélice-boucle-hélice commun à une famille de gènes jouant le rôle de facteurs de transcription (Cubas *et al.*, 1999). Le domaine R est riche en arginine. Les régions exoniques encadrant ces deux domaines sont très variables.

Une portion du gène LEGCYC1A d'environ 520 pb est amplifiée à l'aide des amorces F1 et AR1 (voir Fig. 4.3 ; Citerne, 2005). La solution d'amplification consiste en 10 µl de tampon 5×, 4 µl de $MgCl_2$ à 25 mM, 5 µl de dNTPs à 2 mM/chacun, 3 µl de chaque amorce à 5 µM, 0,25 unité de Promega GoTaq polymerase et 10 à 50 pg d'ADN matrice, dans un volume réactionnel final de 50 µl. Après dénaturation à 94 °C pendant 3 minutes, l'amplification est réalisée en 30 cycles comportant chacun 1 minute à 94 °C, 30 secondes à 50 °C et 1 minute à 72 °C, suivis d'une élongation finale de 5 minutes à 72 °C. Après vérification sur gel, les produits de PCR sont purifiés et quantifiés.

4.5.4 Le gène *SymRK*

Chez les légumineuses, l'interaction avec les bactéries de type Rhizobium et les champignons mycorhiziens nécessite au moins sept gènes dont *SymRK*, NFR1 et NFR5 (Kistner *et al.*, 2005 ; Saito *et al.*, 2007 ; Markmann *et al.*, 2008). Le gène *SymRK* — *Symbiotic Receptor-like Kinase*, aussi connu sous le nom de NORK *(Nodulation Receptor Kinase)*, DMI2 *(Does not make infections 2)* et Ps SYM19 *Pisum sativum Symbiotic*) — est impliqué dans une voie symbiotique commune à la mycorhization et la nodulation (Radutoiu *et al.*, 2003 ; Gherbi *et al.*, 2008). La protéine codée par *SymRK* possède une structure typique d'un récepteur transmembranaire de type kinase, avec un domaine extracellulaire jouant un rôle de récepteur, un domaine transmembranaire, et un domaine kinase interne (voir Fig. 4.4 page suivante). Alors que le domaine kinase est connu pour être généralement bien conservé, le domaine extracellulaire — impliqué dans la reconnaissance de signaux symbiotiques — est plus variable et donc potentiellement plus informatif sur l'évolution du *SymRK*. Nous avons donc ciblé cette région, s'étendant sur environ 3 250 bp chez *Lotus japonicus*.

La première étape a été l'alignement des séquences orthologues de la région extracellulaire du gène *SymRK* des fabacées modèles avec la séquence de *SymRK* de lu-

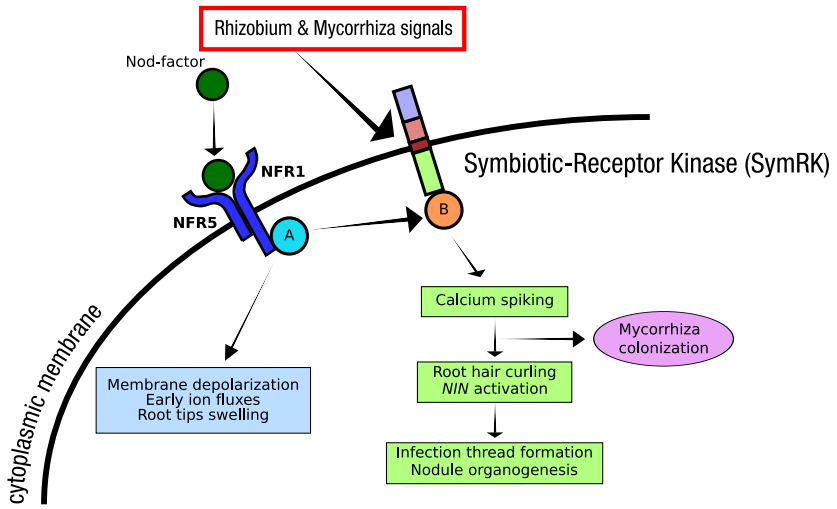

FIG. 4.4 — Schéma et rôle métabolique putatif de la protéine codée par le gène *SymRK* (*Symbiotic Receptor-like Kinase*). La réception d'un signal émis par des bactéries ou des champignons par les protéines codées par les gènes *SymRK* et NFR1-NFR5 entraîne une cascade de changements métaboliques et la formation d'un nouvel organe : le nodule. Adapté de Radutoiu *et al.* (2003).

pin disponible dans GenBank (voir Tab. 4–1). Nous avons ensuite désignés plusieurs amorces dans les zones les mieux conservées en vue d'amplifier et séquencer la région ciblée chez les lupins (voir Mahé *et al.* ; page 104).

TAB. 4–1 — Liste des espèces retenues pour l'analyse comparative des régions codantes du domaine extracellulaire de *SymRK* (a : Endre *et al.*, 2002 ; b : Stracke *et al.*, 2002 ; c : Chen *et al.*, non-publié ; d : Capoen *et al.*, 2005 ; e : Giczey *et al.*, non-publié)

Groupe	**Taxon**	**Nom**	**N° d'accession**
Fabaceae	*Melilotus albus*	MaNORK	AJ428991[a]
	Pisum sativum	SYM19	AJ418376[a]
	Medicago sativa	DMI 2	AJ418368[a]
	Medicago truncatula	MtSYMRK	AF491998[b]
	Astragalus sinicus	AsNORK	AY946203[c]
	Sesbania rostrata	SrSYMRK	AY751547[d]
	Lupinus albus	SymRK	AY935267[e]
	Lotus japonicus	LjSYMRK	AF492655[b]
Fagaceae	*Alnus glutinosa*	AgSYMRK	AY935264[e]
Solanaceae	*Lycopersicon esculentum*	LpSYMRK	AY935266[e]

La solution d'amplification et le programme PCR, identique pour tous les couples d'amorces, consiste en 10 µl de tampon 5×, $MgCl_2$ (2 mM de concentration finale), 5 µl de dNTPs à 2 mM/chacun, 3 µl de chaque amorce à 5 µM, 1,25 unité de Promega GreenTaq, 10 à 50 pg d'ADN matrice, dans un volume réactionnel final de 50 µl. Après dénaturation à 94 °C pendant 150 secondes, l'amplification est réalisée en 30 cycles

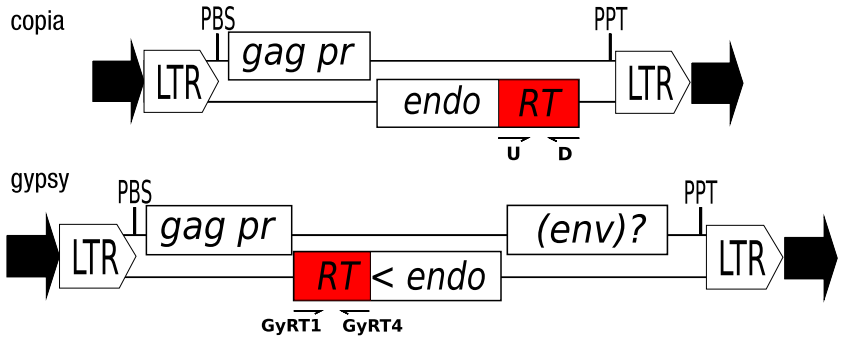

FIG. 4.5 — Localisation de la transcriptase inverse — *reverse transcriptase*, ou *rt* — sur des rétrotransposons Ty1/*copia* (amorces U et D) et Ty3/*gypsy* (amorces GyRT1 et GyRT4), d'après Mhiri & Grandbastien, 2004.

comportant chacun 30 secondes à 94 °C, 1 minute à 54 °C et 90 secondes à 72 °C, suivis d'une élongation finale de 7 minutes à 72 °C. Après vérification sur gel, les produits de PCR sont purifiés, quantifiés et pour certains, clonés.

4.5.5 La transcriptase inverse

Ne disposant pas de ressources génomiques pour les lupins, nous avons utilisé les séquences de transcriptase inverse pour évaluer la diversité des rétrotransposons au sein du genre *Lupinus* et chez plusieurs génistées. Les rétrotransposons, de par leur taille et leurs mécanismes de réplications, sont les éléments transposables les plus susceptibles d'entraîner une augmentation de taille de génome. Les familles Ty1 et Ty3 se sont montrées particulièrement abondantes dans les génomes de plantes, séquencés ou en cours de séquençage — voir par exemple SanMiguel & Bennetzen (1998) ou Piégu *et al.* (2006) —, nous les avons donc ciblées en priorité. Ces rétrotransposons portent une séquence codant pour une transcriptase inverse (*reverse transcriptase* ou *rt*), relativement bien conservée et déjà utilisée par Alix & Heslop-Harrison (2004) dans une démarche similaire.

Les séquences *rt* d'éléments de type Ty1/*copia* et Ty3/*gypsy* (voir Fig. 4.5) ont été amplifiées à l'aide respectivement des amorces U et D (A. J. Flavell *et al.*, 1992), et des amorces GyRT1 et GyRT4 (Friesen *et al.*, 2001). Les amplifiats ont été purifiés, quantifiés puis clonés. La solution d'amplification, identique pour Ty1 et Ty3, consiste en 5 µl de tampon 10×, 4 µl de $MgCl_2$ à 25 mM, 5 µl de dNTPs à 2 mM/chacun, 5 µl de chaque amorce à 10 µM, 2,5 unités de Sigma RedTaq, 10 à 50 pg d'ADN matrice, dans un volume réactionnel final de 50 µl (selon la procédure d'Alix & Heslop-Harrison, 2004).

Ty1/*copia* Après dénaturation à 94 °C pendant 3 minutes, l'amplification est réalisée en 35 cycles comportant chacun 50 secondes à 94 °C, 50 secondes à 39 °C et 1 minute à 72 °C, suivis d'une élongation finale de 7 minutes à 72 °C.

Ty3/*gypsy* Après dénaturation à 94 °C pendant 3 minutes, l'amplification est réalisée en 35 cycles comportant chacun 1 minute à 94 °C, 1 minute à 45 °C et 1 minute à 72 °C, suivis d'une élongation finale de 7 minutes à 72 °C.

4.6 Clonage et purification de plasmides

Pour réaliser le clonage, des bactéries *Escherichia coli* DH5α sont multipliées au laboratoire, stockées et utilisées selon les protocoles décrits par Sambrook & Russell (2001). Le plasmide (aussi appelé vecteur) et la ligase sont fournis par la société Promega : *pGEM-T Vector Systems* et *pGEM-T Easy Vector Systems*. La ligation se fait à 4 °C grâce à l'enzyme T4 (Aslanidis & Jong, 1990), pendant au moins 12 heures. À noter que pour les produits de PCR destinés au clonage, porter le temps d'élongation finale à au moins 60 minutes augmenterait l'efficacité de la ligation (Q.-B. Li & Guy, 1996). Un culot de bactéries DH5α de 50 µl est mélangé à 2 µl de produit de ligation. La transformation se fait par électroporation à 1 800 V suivie d'une mise en culture sur *Lysogeny Broth* (LB) à 37 °C, pendant 2 heures. Entre 50 et 200 µl de milieu de culture sont ensuite étalés sur du LB agar (20 g/l) contenant 10 µg/ml d'ampicilline, 80 µg/ml de 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal) et 120 µg/ml d'isopropyl-β-D-thiogalactopyranoside (IPTG).

Après 24 à 36 heures d'incubation à 37 °C, les colonies blanches sont repiquées sur une nouvelle plaque de LB agar et la présence de l'insert est testée par PCR (amplification utilisant les amorces spécifiques de l'insert). En cas de signal positif, les clones sont mis en culture liquide pendant 12 heures dans environ 5 ml de LB. Les plasmides sont purifiés *via* les kits BioBasic EZ10 *Plasmid DNA* ou Macherey-Nagel *NucleoSpin Plasmid*. À l'issue du clonage, les purifiats sont séquencés à l'aide des amorces T7 et SP6, spécifiques du plasmide (voir *amorces* p. 185).

4.7 Qualité des séquences et déconvolution

Les séquençages ont été réalisés par les sociétés Macrogen (Corée), Agowa (Allemagne) et MWG (France), toutes équipées de séquenceurs capillaires (méthode Sanger). Les séquenceurs ABI 3730xl fournissent en sortie des chromatogrammes que l'algorithme PHRED[17] *(Phil Green revised editor)* (Ewing *et al.*, 1998 ; Ewing & Green, 1998) permet de traduire en séquences nucléotidiques et en fichiers de qualité. Lors de la lecture des chromatogrammes, PHRED attribut une valeur de qualité — sur une échelle théorique allant de 0 à 60, en pratique en-deçà de Q4 le chromatogramme est considéré comme illisible — à chacune des bases de la séquence nucléotidique en la comparant à une prédiction théorique.

$$Q = -10 \times \log_{10}(\text{probabilité d'erreur})$$

17. Le logiciel TRACETUNER fourni le même travail mais semble moins utilisé, voire abandonné. Un nouveau logiciel TRACETOOLS a été publié en 2008. La société Applied Biosystems distibue également le logiciel KB BASECALLER.

Qualité de la base	**Signification**
$0 < x \leqslant 4$	la base est un « N »
10	1 chance sur 10 d'être fausse
20	1 chance sur 100 d'être fausse
30	1 chance sur 1 000 d'être fausse
...	...
60	1 chance sur 1 000 000 d'être fausse

Des valeurs de PHRED sont déduites les longueurs Q16 et Q20. Il s'agit des plus longues portions de séquences dont les bases ont une valeur de PHRED supérieures à 16 ou à 20 (respectivement 97,5 % et 99 % de confiance)[18].

Si les produits séquencés sont hétérogènes en taille (insertion–délétion), la lecture des chromatogrammes peut être très difficile. En effet, un décalage d'une base (ou plus) entraîne un doublement de tous les pics dans le reste du chromatogramme (voir Fig. 4.6) et rend la séquence illisible. Récemment, des algorithmes dédiés à la résolution de ce problème ont été publiés. Schématiquement, le programme tente à partir d'un *base calling* complet basé sur le code IUPAC[19] de « deviner » la taille et la nature de l'insertion pour pouvoir reconstituer les séquences d'origine. C'est sur ce principe qu'est basé INDELLIGENT (Dmitriev & Rakitov, 2008)[20].

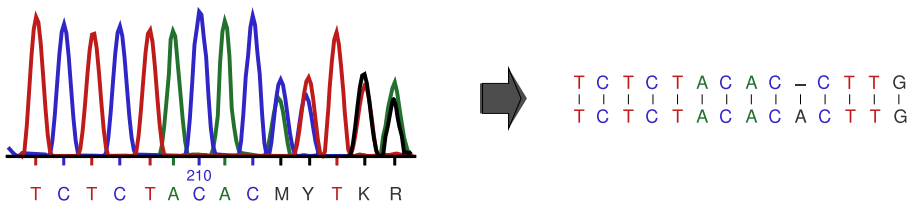

FIG. 4.6 — Exemple de chromatogramme au format AB1 avec sa traduction en séquence nucléotidique. À partir de la 213e base, le signal est brouillé et le *base calling* se fait sur la base du code IUPAC. Les outils de déconvolution comme INDELLIGENT (Dmitriev & Rakitov, 2008), permettent de révéler le signal sous-jacent — ici l'insertion d'un A entraînant un décalage — et d'en déduire les séquences présentes dans le mélange.

Dans sa version courante, le logiciel PHRED ne permet pas d'obtenir des séquences au format IUPAC. Cependant, j'ai pu obtenir d'Ewing & Green une version bêta capable de réaliser un *base calling* complet. Nous avons ainsi pu tester INDELLIGENT et valider cette démarche de déconvolution sur certaines séquences n'ayant pu être obtenues en séquençage direct et pour lesquelles une étape de clonage avait été nécessaire. La généralisation de cette technique devrait permettre d'améliorer significativement la qualité du séquençage Sanger et d'éviter un recours systématique au clonage.

18. Pour mémoire la probabilité d'erreur p peut être retrouvée par $p = 10^{-\frac{Q}{10}}$.
19. Nomenclature de l'*International union of pure and applied chemistry*, voir annexe B page 184.
20. `http://ctap.inhs.uiuc.edu/dmitriev/indel.asp`

4.8 Filtrage et vérification de la nature des séquences

Cette étape, à la fois primordiale et gourmande en ressources humaines, a été automatisée dans le cadre des premiers grands projets de séquençage. Le logiciel LUCY (Chou & Holmes, 2001 ; S. Li & Chou, 2004) se base sur la lecture de chromatogrammes et sur les valeurs de PHRED pour éliminer les portions de séquences dont la lecture est douteuse. Les portions restantes sont ensuite comparées à une banque de séquences de vecteurs pour éliminer d'éventuelles contaminations. À l'issue du processus de filtrage, jusqu'à 50 % du nombre total de bases peut-être rejeté.

Une fois les séquences nettoyées, il s'agit de vérifier qu'elles correspondent à la portion de génome ciblée. Pour ce faire, les séquences sont soumises à BLAST[21] (Altschul *et al.*, 1990) : variantes BLASTN ou BLASTX en fonction du degré de divergence des séquences traitées. Plusieurs informations peuvent en être tirées : le pourcentage d'identité entre la séquence soumise et la séquence-cible, la longueur de recouvrement et surtout, l'orientation de la séquence soumise par rapport à la séquence de référence (plus/moins). Ce dernier point permet de repérer et de traiter les séquences à inverser-complémenter[22].

Pour les régions de grande taille, comme le *SymRK*, il est nécessaire d'amplifier et de séquencer le gène par fragments (*walking PCR*), ces fragments sont ensuite assemblés par CAP3 (Huang & Madan, 1999). Avant de procéder à l'alignement des séquences obtenues, il est possible d'éliminer les *outliers*, c'est-à-dire les séquences présentant moins de 40 % d'identité avec les autres séquences de la matrice. En effet, en-deçà de ce seuil, l'alignement multiple est fortement perturbé et peut conduire à de fausses interprétations. Ces différentes étapes de préparation de séquences ont été automatisées et intégrées dans une série de scripts (disponibles sur demande).

---

21. Il existe des alternatives à BLAST comme par exemple FASH (Veksler-Lublinksy *et al.*, 2008) basé sur l'algorithme de la « transformée de Fourier rapide ».

22. Étape au cours de laquelle la séquence est inversée (5′-ATGC-3′ devient 3′-CGTA-5′) et complémentée (5′-CGTA-3′ devient 5′-GCAT-3′).

5

Analyse phylogénétique et annotation de séquences

« *All models are wrong but some are useful.* »
George Box (1976)

L'essor de la biologie moléculaire est lié à celui de l'outil informatique. Dans les années 1970, les premiers programmes cherchant à modéliser l'évolution moléculaire ont fait leur apparition, et la démocratisation dans les années 1980 des micro-ordinateurs a ensuite considérablement accéléré le processus. Aujourd'hui, les outils bioinformatiques se diversifient et se complexifient pour faire face à la croissance exponentielle du corpus de données moléculaires. La méthodologie utilisée au cours de ce travail de thèse repose principalement sur deux axes : l'analyse phylogénétique et l'annotation de séquences. Les stratégies et les outils utilisés pour ces deux axes de travail seront décrit dans ce chapitre.

Retour aux fondamentaux

L'alignement de séquences est une des opérations les plus courantes en bioinformatique. Son importance est fondamentale puisque c'est sur elle que se basent beaucoup de résultats comme les mesures d'identités et la recherche d'homologies entre séquences. Pour chaque alignement de n séquences (l'alignement de deux séquences étant un cas particulier d'alignement multiple), il existe une solution optimale. Pour être certain d'avoir atteint cette solution, il n'y a pas d'autres alternatives que l'évaluation de toutes les solutions possibles. Au-delà de quelques dizaines de séquences, cette énumération systématique devient une impossibilité physique, le temps et l'espace mémoire nécessaires dépassant les dimensions caractéristiques de notre Univers[1].

1. L'accroissement des performances des ordinateurs ne rendra pas le problème plus accessible : « que vous utilisiez une petite cuillère ou un seau pour vider l'océan, l'ampleur de la tâche ne change guère ». Pour une approche plus formelle de cette problématique, on consultera *Calculateurs, calculs et calculabilité* de Ridoux & Lesventes (2008).

Le problème de la reconstruction phylogénétique est similaire. Selon Cavalli-Sforza & Edwards (1967), cités par Felsenstein (1978b), le nombre d'arbres dichotomiques non-enracinés pour un jeu de n taxons se calcule de la façon suivante[2] :

$$N_{arbres} = \frac{(2n-5)!}{2^{n-3}(n-3)!} \quad \text{ou} \quad N_{arbres} = \prod_{k=3}^{n}(2k-5)$$

Pour $n = 20$, le nombre d'arbres non-enracinés est supérieur à 8×10^{21}, largement hors de portée des outils informatiques actuels. L'alignement multiple et la reconstruction phylogénétique relèvent de l'optimisation combinatoire, une classe de problèmes dits « NP-complets » (L. Wang & Jiang, 1994), particulièrement difficiles à résoudre de façon exacte et pour lesquels il a fallu développer des méthodes approchées dites « heuristiques » (Roch, 2006 ; Z. Zhang *et al.*, 2008).

Tous les algorithmes développés aujourd'hui se fondent donc sur des heuristiques permettant d'approcher en un temps raisonnable la solution optimale, mais sans certitude de l'avoir atteinte. Par analogie, et dans l'optique d'une recherche d'optimum, l'espace des solutions peut être assimilé à un paysage accidenté dans lequel il s'agit de localiser les pics les plus hauts (ou les vallées les plus profondes). Pour accélérer l'exploration de cet espace gigantesque, plusieurs algorithmes ont été mis au point (liste non-exhaustive) : procédure par séparation et évaluation (*branch & bound*), algorithme glouton, méthodes de voisinage, méthode de recherche « tabou », méthode de « descente stochastique » (recuit simulé, Monte-Carlo) et algorithmes génétiques.

Ces méthodes, qui peuvent être associées entre elles, fournissent en un temps raisonnable une solution proche de la solution optimale. La recherche peut-être interrompue à tout moment et fournit toujours une réponse dont la précision croît en fonction du temps écoulé. C'est le type et l'intelligence des heuristiques utilisées qui distingue les différents logiciels, leur rapidité et leur précision.

5.1 Reconstruction phylogénétique

Dans cette section, nous décrirons notre premier axe méthodologique concernant l'analyse phylogénétique. En construisant des phylogénies de séquences, notre objectif est double. Dans un premier temps, il s'agit de clarifier les relations de parenté existant entre les différentes lignées de lupins, avec comme hypothèse sous-jacente l'idée que la phylogénie des gènes reflète celle des organismes. Dans un deuxième temps, en appliquant les techniques de reconstruction phylogénétique à des séquences d'éléments transposables, nous avons cherché à évaluer leur diversité et leur fréquence au sein de certaines lignées de lupins. Une vue générale de la démarche suivie est donnée dans la Fig. 5.1 page suivante.

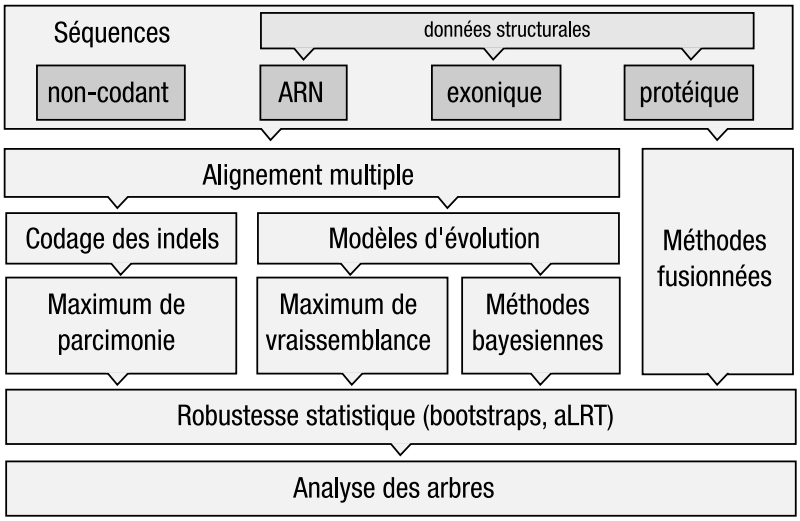

FIG. 5.1 — Description de la méthodologie employée pour les analyses phylogénétiques présentées dans ce travail de thèse. L'alignement multiple se fait en fonction du rôle des séquences. L'alignement des séquences codantes doit se baser autant que possible sur les informations disponibles pour les niveaux supérieurs : données structurales pour les ARN ayant une activité enzymatique (ARN polymérase, ARNr, ARNt, ARNtm), séquences protéiques et données structurales pour les séquences traduites. Les « méthodes fusionnées » désignent les nouvelles stratégies réalisant en une seule étape l'optimisation de l'alignement multiple, des paramètres d'évolution et de la topologie de l'arbre.

5.1.1 Alignement multiple

Établir l'homologie entre les bases de plusieurs séquences, c'est-à-dire les aligner, est une étape critique dans le processus de reconstruction phylogénétique (Morrison & Ellis, 1997). En effet, tout le reste de la chaîne de traitement va être affectée par les choix faits à cette étape. Or, il est courant de lire qu'après un alignement logiciel la matrice a été retouchée « à la main ». Cette pratique, si elle était parfaitement défendable il y a quelques années, ne l'est plus aujourd'hui. Corriger manuellement une matrice, peut introduire des biais liés, entre autres, au fonctionnement de notre cerveau — notamment notre goût pour la symétrie — mais surtout, c'est perdre un caractère essentiel de la preuve scientifique, la répétabilité. Les algorithmes actuels ne sont pas dénués de biais, mais ceux-ci ont l'avantage d'être connus et paramétrables[3]. On peut citer le biais d'asymétrie 5'–3'/3'–5' (States & Boguski, 1991 ; Martin *et al.*, 2007 ; Landan & Graur, 2007) et le biais d'asymétrie de traitement des insertions et des délétions (les insertions sont plus lourdement pénalisées que les délétions). Ce dernier artefact devient dominant quand la distance évolutive entre les séquences augmente et conduit à des

2. Pour mémoire, $n! = \prod_{k=0}^{n-1}(n-k)$.

3. Pour estimer les performances de ces algorithmes, ceux-ci sont confrontés à des banques de jeux de séquences réelles (Whelan *et al.*, 2006) ou obtenues par simulation (Varadarajan *et al.*, 2008). C'est la nature de ces banques de test qui peut poser problème ainsi que le risque de sur-apprentissage des algorithmes.

alignements artificiellement denses et donc à l'alignement de régions non-homologues (Landan & Graur, 2009). Une solution algorithmique existe ; elle passe par l'évaluation des implications phylogénétiques de chaque création d'insertions-délétions (*indels*) tout au long du processus d'alignement multiple (Loÿtynoja & Goldman, 2005, 2008a, 2008b). À terme, cela implique la fusion de l'étape d'alignement et de l'étape de reconstruction phylogénétique en une seule opération (voir section 5.1.3.5 page 86).

La courte histoire des algorithmes d'alignement multiple est marquée par une domination écrasante du logiciel CLUSTAL (J. D. Thompson *et al.*, 1994 ; Chenna *et al.*, 2003). Pourtant, depuis sa publication en 1994, CLUSTAL n'a plus reçu d'améliorations significatives (Edgar & Batzoglou, 2006), tandis que plus de 50 nouvelles méthodes d'alignement multiple ont été décrites (Wallace *et al.*, 2006). Pourquoi un tel foisonnement ? Si l'alignement d'une paire de séquences est un problème résolu, l'alignement multiple présente un problème beaucoup plus complexe, impossible à résoudre de façon optimale pour des jeux de données dépassant quelques dizaines de séquences (Batzoglou, 2005 ; Morrison, 2006 ; Roch, 2006). En réponse à ce problème, deux grandes familles de méthodes ont été développées, l'alignement global et l'alignement local.

CLUSTAL fait partie des méthodes dites d'« alignement global » (Needleman & Wunsch, 1970) et est d'une remarquable efficacité tant que le jeu de données ne présente pas de complexité particulière. Les données traitées sont représentées sous forme de matrice et l'algorithme fournit rapidement une vue générale de la proximité des séquences. Par contre, si les séquences présentent d'importantes différences — longs indels, régions hypervariables, portions non-homologues —, l'algorithme va forcer leur alignement et de fait, introduire de l'homoplasie et faire baisser le ratio signal/bruit.

Les méthodes dites d'« alignement local » (T. F. Smith & Waterman, 1981) corrigent ce défaut. Elles permettent de limiter l'alignement à des zones jugées pertinentes. Pour gagner en qualité de signal, une partie de l'information est donc négligée. En pratique, cette méthode permet de traiter des séquences plus éloignées phylogénétiquement que ce qui peut être traité par CLUSTAL. Les algorithmes d'alignements locaux les plus cités pour leur efficacité sont PROBCONS (Do *et al.*, 2005), MUSCLE (Edgar, 2004), MAFFT[4] (Katoh *et al.*, 2002, 2005 ; Katoh & Toh, 2007, 2008 ; Katoh *et al.*, 2009), KALIGN2 (Lassmann *et al.*, 2009), DIALIGN (Morgenstern *et al.*, 1998 ; Morgenstern, 1999 ; Schmollinger *et al.*, 2004 ; Subramanian *et al.*, 2005) et T-COFFEE (Notredame *et al.*, 2000).

Comme pour d'autres domaines de la bioinformatique — par exemple la prédiction de structure secondaire (Cuff *et al.*, 1998) ou la prédiction de gènes (Allen & Salzberg, 2005) —, il a été montré que les méta-méthodes fournissent souvent des résultats meilleurs que les différentes méthodes prises indépendamment. C'est la voie choisie par Wallace *et al.* (2006) avec le logiciel M-COFFEE. Le jeu de séquences est soumis à douze algorithmes parmi les plus performants, et les différents alignements obtenus sont fusionnés en un seul alignement consensuel. C'est l'outil que nous avons retenu pour nos analyses. Cependant, pour certaines matrices de grande taille comme les ma-

4. `http://align.bmr.kyushu-u.ac.jp/mafft/software/`

trices de *reverse transcriptase*, nous nous sommes tournés vers MAFFT, moins gourmand en mémoire.

Pour les séquences transcrites mais non-traduites comme les ITS, les données structurales ont été utilisées selon la méthode décrite par Schultz & Wolf (2009). Pour les séquences traduites, il est nécessaire de réaliser l'alignement sur des séquences d'acides aminés et, si possible, d'y intégrer les données structurales pour optimiser l'alignement nucléotidique. Ces étapes peuvent être réalisées *via* les logiciels MAGNOLIA (Fontaine *et al.*, 2008) ou T-COFFEE et ses variantes EXPRESSO et RCOFFEE. En pratique, le *rbc*L est le seul gène utilisé au cours de nos analyses pour lequel suffisamment de données structurales sont disponibles pour guider l'alignement (Andersson & Backlund, 2008). Au niveau infra-générique auquel nous nous situons, l'apport de ces données supplémentaires est faible voire nul. Cependant, elles ont été utilisées systématiquement, par souci de rigueur et de cohérence.

Perspectives pour l'alignement multiple Les logiciels actuels représentent les séquences sous forme de matrice. Or, les séquences peuvent être le résultat de recombinaisons, de duplications ou d'inversions. De nouvelles approches basées sur la représentation par des graphes permettent d'intégrer ce phénomène de modularité des séquences. En effet, la représentation sous forme de graphe est plus riche et expressive que la représentation matricielle. On peut citer, par ordre d'expressivité croissant : POA *(Partial Order Alignment)* (C. Lee *et al.*, 2002), ABA *(A-Bruijn Alignment)* (Raphael *et al.*, 2004) et PROTOMATA[5] (Kerbellec, 2008).

Une autre possibilité est de contourner le problème en ne faisant pas d'alignement. De nouvelles méthodes d'inférence phylogénétiques sont apparues récemment. Elles ont pour principal avantage d'accepter des jeux de données plus grands et plus complexes (renversements, duplications et translocations) que les méthodes classiques et de ne pas nécessiter de phase d'alignement (Ferragina *et al.*, 2007). Ces méthodes se répartissent en trois catégories (Lu *et al.*, 2008) : 1) contenu en gènes *(gene contents)* ; 2) compression de données *(data compression)* et 3) composition de chaîne *(string composition)*. C'est cette dernière qui a reçu le plus d'attention de la part de la communauté scientifique (voir par exemple Apostolico & Denas, 2008).

À l'issue de cette étape d'alignement multiple, le jeux de données est prêt pour l'étape de reconstruction phylogénétique. Si plusieurs approches sont possibles, trois grandes familles de méthodes sont généralement citées : la méthode phénétique basées sur des mesures de distance entre séquences (méthode non-décrite ici), les méthodes paramétriques basées explicitement sur des modèles d'évolution, et la méthode cladiste, définie comme non-paramétrique, puisque ne faisant pas explicitement appel à un modèle d'évolution.

5. `http://protomata-learner.genouest.org/`

5.1.2 Méthodes non-paramétriques

Les principes de l'analyse cladistique[6] ont été élaborés par l'entomologiste Willi Hennig dans son ouvrage de taxinomie fondamentale (1950), ainsi que dans sa synthèse sur la phylogénie des insectes (1969). Dans le système cladistique, la phylogénie est reconstruite à l'aide d'une analyse de caractères qui vise à identifier les états plésiomorphes (primitifs) et apomorphes (dérivés). Les parentés entre les taxons étudiés sont identifiées sur la base des seuls états apomorphes partagés par tel et tel taxon, ce que l'on appelle les synapomorphies. Les synapomorphies sont imputées à un héritage à partir d'une espèce ancestrale propre aux taxons qui les possèdent. Les groupes ainsi construits sont monophylétiques.

Pour maximiser le contenu informationnel d'une matrice de séquences, il convient de coder les insertions-délétions, normalement considérées comme de l'information manquante par les logiciels de reconstruction phylogénétique. En effet, celles-ci peuvent être partagées par un sous-ensemble des taxons analysés et donc être synapomorphiques.

Codage des insertions-délétions Ogden & Rosenberg (2007) ont réalisé un travail de simulation destiné à évaluer l'intérêt du codage des indels. Pour se faire, ils ont aligné et analysé, à l'aide de CLUSTAL et de PAUP (Swofford, 1989-2003), 15 400 topologies obtenues avec codage des indels (*fith-state* ou *simple indel coding*) et sans codage (c'est-à-dire indels traités comme des données manquantes). Leurs résultats montrent que, sur les matrices testées, dans 82 % des cas les topologies obtenues avec codage ou sans codage sont identiques. Pour les 18 % restants, dans 90 % des cas c'est la matrice codée qui donne le meilleur résultat (voir Fig. 5.2). Pour des matrices de séquences très divergentes et des topologies très déséquilibrées (*pectinate tree shape*), le codage des indels peut avoir un impact négatif. Cependant, Ogden & Rosenberg prônent l'utilisation systématique du codage des indels et soulignent les bénéfices futurs attendus de l'amélioration des méthodes d'alignement.

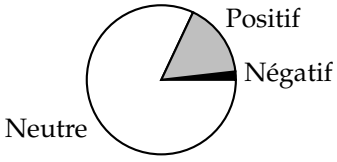

FIG. 5.2 — Effet du codage des insertions-delétions sur la qualité de la reconstruction phylogénétique (d'après Ogden & Rosenberg, 2007) : dans 82 % des cas il n'y a pas de différence de résultats entre matrice codée et non-codée. Le codage a un effet positif dans 16.2 % des cas et négatif pour les 1.8 % restants. Les alignements ont été réalisés par CLUSTAL.

6. Hennig n'a pas utilisé dans ses différents ouvrages les termes « cladisme », « analyse cladistique », « cladogramme », ou tout simplement « clade », tous dérivés de la racine grecque κλαδος (branche). Le cladisme (ou la cladistique) y est dénommée « systématique phylogénétique », le cladogramme est un « schéma d'argumentation phylogénétique », le clade est un « groupe monophylétique » (Darlu & Tassy, 1993).

Quelle méthode de codage utiliser ? C'est la méthode de Véronique Barriel (1994) qui a été utilisée jusqu'à présent par notre équipe. Cependant, Freudenstein & Chase (2001) préconisent l'usage de caractères multi-états non-ordonnés (*unordered multistate characters*) excluant de fait les méthodes basées, comme celle de Barriel, sur un encodage ternaire (présence, absence, inconnu). C'est la méthode de Simmons & Ochoterena (2000) — *modified complex indel coding* (MCIC) — qui se montre la plus efficace mais également la plus complexe à mettre en œuvre (Simmons *et al.*, 2007). De fait, elle a principalement été utilisée dans sa version dite « simple », jusqu'à ce que Müller (2005b, 2006) intègre à son logiciel SEQSTATE la version dite « complexe » des règles établies par Simmons & Ochoterena. C'est donc la méthode de codage MCIC qui a été utilisée pour nos analyses en maximum de parcimonie[7].

Maximum de parcimonie La recherche de l'arbre le plus parcimonieux peut être réalisé par plusieurs logiciels, le plus populaire étant PAUP (Swofford, 1989-2003). Nous l'avons utilisé avec les paramètres suivants : recherche heuristique *random stepwise addition* et *tree bisection-reconnection* (TBR)[8], les branches de longueur nulle sont effondrées, tous les caractères ont le même poids et les gaps sont traités comme des données manquantes (ce qui est compensé par le codage des indels réalisé en amont). La méthode du maximum de parcimonie présente l'avantage d'être plus rapide que les méthodes paramétriques (et d'être basée sur un corpus d'hypothèses plus réduit). Cependant, elle peut également être plus sensible au phénomène d'attraction de longues branches (Felsenstein, 1978a ; Siddall & Whiting, 1999 ; Pol & Siddall, 2001).

Ce phénomène s'explique de la façon suivante : du fait de la pauvreté de l'alphabet nucléotidique (A, C, G et T), deux séquences éloignées phylogénétiquement mais évoluant rapidement peuvent accumuler des substitutions convergentes. Interprétées à tort comme des synapomorphies, ces mutations provoquent le regroupement artificiel de ces séquences (Bergsten, 2005). Afin de détecter cet artefact, nous avons utilisé systématiquement les méthodes paramétriques en parallèle de la méthode du maximum de parcimonie.

Bootstrapping Le *bootstrap* est une technique statistique utilisée pour évaluer la robustesse d'une reconstruction phylogénétique (Felsenstein, 1985b). Il est souvent considéré à tort comme la probabilité qu'un clade soit réel, alors qu'il devrait être au mieux considéré comme un indicateur du degré de support d'une technique particulière, pour un clade particulier et pour un jeu de données particulier (Hillis & Bull, 1993). D'autres interprétations ont été proposées, voir Z. Yang & Rannala (2005) pour revue. Un *bootstrap* est un tirage avec remise des caractères de la matrice de départ permettant de constituer de nouvelles matrices de taille identique. En d'autres termes, ce test permet d'évaluer la résistance de la reconstruction suite à une perturbation du

7. Les méthodes paramétriques actuelles (maximum de vraisemblance et analyse bayésienne), traitent les gaps comme des caractères indéterminés (pour revue, Dwivedi & Gadagkar, 2009).

8. Müller (2005a) a démontré que les opérations autres que le *tree bisection-reconnection* augmentent le temps de calcul sans apporter de gain notable.

jeu de données initiales par pondération aléatoire des différents caractères. Chacune de ces nouvelles matrices est soumise à une recherche de l'arbre le plus parcimonieux. Ces arbres sont ensuite combinés en un seul arbre de consensus majoritaire. Les clades apparaissant dans plus de 50 % des cas recevant alors une valeur de *bootstrap* comprise entre 50 et 100 % (un soutien supérieur à 70 % est considéré comme bon selon Hillis & Bull, 1993).

Une valeur de bootstrap représente une répartition observée entre deux états (présence ou absence du clade). Cette répartition observée est une approximation plus ou moins précise de la vraie répartition (lorsque le nombre de réplicats tend vers l'infini). En pratique, combien doit-on réaliser de réplicats ? La réponse dépend de la précision souhaitée. La figure 5.3 montre clairement que la précision des valeurs de bootstrap croît avec le nombre de réplicats mais également avec le déséquilibre de répartition : l'incertitude est faible pour un clade présent dans 99 % des cas, même avec un faible nombre de réplicats. Le seuil de 1 000 réplicats est un bon compromis entre précision et temps de calcul, c'est le seuil que nous nous sommes fixés au cours de ce travail de thèse, tant pour le maximum de parcimonie que pour les méthodes paramétriques.

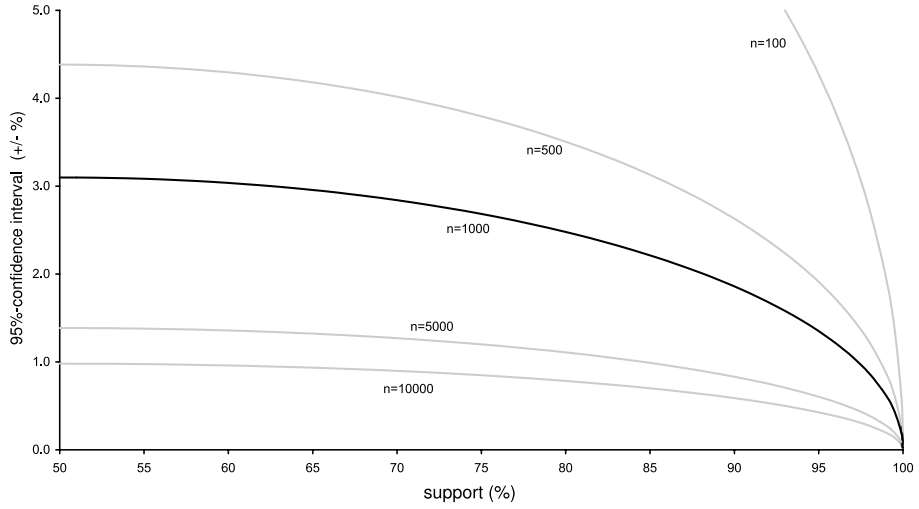

FIG. 5.3 — Relation entre le nombre de réplicats et la précision des valeurs de bootstrap : intervalle de confiance calculé en pourcentage pour un risque de première espèce fixé à 5 %, (d'après Müller, 2005a). Par exemple, pour une valeur de bootstrap observée de 70 % après 1 000 réplicats, la valeur réelle a 95 % de chances d'être comprise entre 67 et 73 %.

5.1.3 Méthodes paramétriques

Les méthodes non-paramétriques, comme le maximum de parcimonie, sont concurrencées par les méthodes paramétriques — c'est-à-dire utilisant des modèles dont les paramètres peuvent être modifiés — aussi désignées « méthodes probabilistes ». L'émergence de ces méthodes probabilistes, en parallèle avec l'informatisation et l'arrivée des données moléculaires, ne s'est pas faite sans heurts. En 2001, Joseph Felsen-

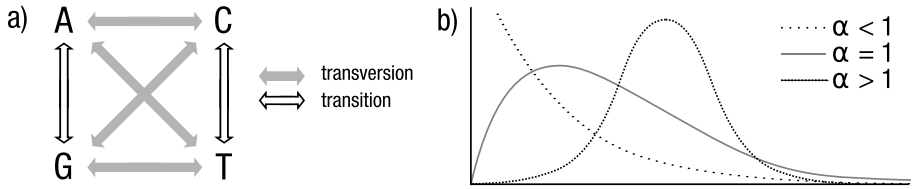

FIG. 5.4 — Transition, transversion et forme de la distribution gamma (Γ). Les modèles d'évolution moléculaire font des hypothèses sur les taux de substitution d'une base par une autre. a) On distingue deux types de substitutions : les transitions (A ⇔ G et C ⇔ T) et les transversions (A ⇔ C, A ⇔ T, C ⇔ G et G ⇔ T). À chacun de ces taux estimés lors de l'analyse paramétrique, peut être appliqué la courbe de distribution Γ. b) Cette distribution autorise une variation des taux autour de la valeur estimée. Le paramètre α décrit la forme de la distribution Γ : si $\alpha < 1$, la distribution Γ est écrasée (plus d'hétérogénéité), si à l'inverse $\alpha > 1$, la distribution Γ se rapproche d'une distribution normale (plus d'homogénéité).

stein a décrit la période 1960-2000 et les schismes qui se sont créés entre les partisans d'une cladistique pure et le reste de la communauté des phylogénéticiens (voir également Siddall & Whiting, 1999). Aujourd'hui, ces méthodes connaissent un succès croissant et ce sont elles qui atteignent les meilleurs résultats sur les jeux de tests.

5.1.3.1 Modèles d'évolution

Pour tirer le meilleur parti de ces méthodes, il est important de sélectionner le modèle d'évolution le plus approprié pour le jeu de données. Un modèle est une matrice de nombres représentant des taux de substitution, par exemple le remplacement d'un A par un T, ou celui d'une proline par une arginine. En fonction de la nature de la séquence, ces matrices décrivent les substitutions entre 4 (nucléotides), 20 (acides aminés) ou 61 états (codons[9]) pour lesquels les valeurs ont été fixées à partir de l'observation de grands corpus de données. Du point de vue du modèle, ces processus de mutation s'appliquent de manière identique tout au long de la séquence. Or, il existe des contraintes structurales et fonctionnelles qui donnent plus d'importance à certains sites et moins à d'autres. Pour prendre en compte ces contraintes et améliorer le réalisme des modèles, les fonctions correctrices I, Γ et F ont donc été introduites (Whelan, 2008).

La fonction I (Reeves, 1992), de valeur comprise entre 0 et 1, décrit la proportion de sites invariants. Il ne s'agit pas du nombre de sites invariants observés dans la matrice mais d'une estimation du nombre de sites n'accumulant pas de mutations. Si I est fixé à zéro, tous les sites sont considérés comme susceptibles de muter au cours du temps.

La fonction Γ (Z. Yang, 1993) autorise une dispersion des taux de transition et de transversion autour d'une valeur estimée. Le paramètre α décrit la forme de la distribution Γ estimée lors de la recherche du meilleur modèle. Si $\alpha < 1$, la distribution Γ adopte une forme en L, avec un petit nombre de sites évoluant rapidement, une majorité de sites conservés, et des taux de transition/transversion plus susceptibles d'être

9. Les trois codons-stops UAA, UAG et UGA sont exclus.

éloignés de la valeur estimée (hétérogénéité). Si à l'inverse $\alpha > 1$, la distribution Γ se rapproche d'une distribution normale[10] avec moins de variation des taux d'un site à l'autre (taux homogènes)[11].

Les modèles d'évolution peuvent également inclure des données sur la fréquence des nucléotides ou des acides aminés. La fonction F (Cao *et al.*, 1994) permet de prendre en considération les fréquences observées dans la matrice soumise à analyse.

De nouveaux modèles, plus complexes et donc plus réalistes sont développés activement[12]. Citons par exemple les *codon-position models* (Shapiro *et al.*, 2006 ; Bofkin & Goldman, 2007 ; Rodrigue *et al.*, 2009), les *mixture models* (Le *et al.*, 2008) ou le concept d'hétérotachie (Pagel & Meade, 2008 ; H.-C. Wang *et al.*, 2009). L'introduction du concept général d'hétérotachie (méthode des *concomitantly variable codons* ou *covarions*) autorise des variations des taux de substitution dans le temps, là où la fonction Γ permet uniquement des variations des taux de substitution le long de la séquence. La méthode des covarions, puisqu'elle permet de décrire des rythmes d'évolutions variables d'une branche à l'autre, diminue l'impact du phénomène d'attraction des longues branches.

Qu'est-ce que la vraisemblance et comment choisit-on un modèle ? La vraisemblance L (*likelihood*) est proportionnelle à la probabilité d'observer le jeu de données D connaissant le modèle d'évolution M, le vecteur des K paramètres du modèle θ, la topologie de l'arbre τ et le vecteur représentant les longueurs de branches v (le centre d'intérêt étant dans ce cas le modèle, ce sont ses paramètres qui sont optimisés). L'analyse renvoie une valeur synthétisant l'accord entre le modèle et les données : le logarithme de la vraisemblance ℓ (*log-likelihood*).

$$L = P(D|M, \theta, \tau, v) \quad \Longrightarrow \quad \ell = \ln P(D|M, \hat{\theta}, \hat{\tau}, \hat{v})$$

Pour pouvoir décider du modèle le plus approprié — c'est-à-dire présentant le meilleur équilibre entre *réalisme* et *coût calculatoire* —, il faut disposer d'un critère d'évaluation (Ripplinger & Sullivan, 2008, pour revue). Il existe plusieurs types de tests : *hierarchical likelihood ratio tests* (hLRT ; Goldman, 1993), *dynamical likelihood ratio tests* (dLRT), *Akaike information criterion* (AIC ; Akaike, 1974), *Bayesian information criterion* (BIC) et *decision theory method* (DT). Les travaux de Posada & Buckley (2004) montrent que c'est la stratégie AIC qui s'avère la plus pertinente, nous l'avons donc retenue pour nos analyses. L'AIC peut être considéré comme la mesure de la quantité

10. La distribution normale est un cas particulier de distribution Γ.

11. Pour simplifier les calculs, la distribution Γ est discrétisée en n classes, c'est à dire qu'elle peut prendre n valeurs. Il est facile d'imaginer qu'un plus grand nombre de classes rendra le modèle plus réaliste et améliorera la vraisemblance de l'analyse, cependant la littérature ne donne que peu d'informations sur la valeur optimale à adopter. Les tests réalisés sur nos jeux de données avec 4, 10, 20 et 40 classes montrent bien un gain significatif de vraisemblance mais au prix d'un temps de calcul décuplé. De plus, Galtier & Jean-Marie (2004) ont montrés que les effets d'une augmentation du nombre de classes sont d'autant plus grands que le nombre de séquences augmente. Ce manque de recul nous a poussé à laisser ce paramètre sur sa valeur par défaut de 4 classes (ou Γ_4, écrit Γ dans le reste du manuscrit).

12. En contrepartie, la complexification des modèles augmente le risque de tomber dans un surapprentissage des jeux de tests et diminue leur capacité prédictive (Steel, 2005).

d'information perdue du fait de l'utilisation d'un modèle pour approximer le processus d'évolution moléculaire. L'AIC le plus petit est donc le meilleur. En prenant K le nombre de paramètres du modèle (ce qui inclut les longueurs de branches), on obtient :

$$\mathrm{AIC} = -2\ell + 2K$$

Si n, la taille du jeu de données[13] est faible (par exemple $n/K < 40$), il faut introduire un facteur correctif et calculer l'AIC_c (Sugiura, 1978) :

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2K(K+1)}{n-K-1}$$

Le facteur correctif tend vers zéro quand la taille du jeu de données devient très grande. Par conséquent, l'AIC_c a été utilisé pour tous les jeux de données traités au cours de ce travail, comme suggéré par Posada & Buckley (2004).

La hiérarchisation des modèles se calcule simplement en mesurant la différence Δ_i entre le meilleur AIC et l'AIC du modèle i. Ces différences permettent de calculer w_i, le poids relatif de chaque modèle, qui peut être interprété comme la probabilité que le modèle soit la meilleure approximation étant donné la matrice de séquences[14].

$$\Delta_i = \mathrm{AIC}_i - \min(\mathrm{AIC}) \quad \text{et} \quad w_i = \frac{\mathrm{e}^{-\frac{1}{2}\Delta_i}}{\sum_{r=1}^{R} \mathrm{e}^{-\frac{1}{2}\Delta_r}}$$

La somme des poids w_i des différents modèles étant égale à 1, il est possible de fixer un seuil arbitraire (intervalle de confiance) et de ne sélectionner que les modèles représentant 80, 90 ou 95 % du poids total. Un consensus des topologies estimées par chacun des modèles — ou pour les modèles inclus dans l'intervalle de confiance désigné — peut être construit et pondéré par les valeurs w calculées précédemment. Cette opération permet d'obtenir une topologie avec une valeur de support pour chaque branche, valeur correspondant à la sensibilité face à un changement de modèle. Un support de 1 indique que la branche n'est pas sensible au type de modèle utilisé ; un support de 0,4 par exemple indique que cette branche n'est supportée que par certains modèles.

Le logiciel JMODELTEST (Posada, 2008, 2009), une évolution récente de MODELTEST (Posada & Crandall, 1998 ; Posada, 2006), permet de tester 88 modèles différents sur des séquences nucléotidiques. Pour les séquences protéiques, la sélection du meilleur modèle d'évolution se fait *via* PROTTEST (Abascal *et al.*, 2005). Ce sont ces deux logiciels, JMODELTEST et PROTTEST, qui ont été utilisés pour les analyses décrites dans les chapitres 6 et 7.

13. Cette notion est floue dans le cas d'un alignement de séquences nucléotidiques. Plusieurs paramètres sont utilisés — longueur de l'alignement, nombres de sites variables (entropie), nombre de séquences, *etc* — sans que la question ait été tranchée (Posada, 2009).

14. Des modèles présentant une faible différence ($\Delta_i < 2$) avec le meilleur modèle peuvent être pris en compte. Si par contre $\Delta_i > 10$, le modèle n'est pas supporté par les données et peut être rejeté.

5.1.3.2 Maximum de vraisemblance

La méthode du maximum de vraisemblance rencontre un succès important et est utilisée dans un grand nombre de travaux phylogénétiques. Il existe de nombreuses implémentations de cette méthode mais le logiciel PHYML (Guindon & Gascuel, 2003 ; Guindon *et al.*, 2005, 2009) est parmi ceux ayant prouvé leur efficacité, leur rapidité et la pertinence de leurs résultats. Nous avons donc choisi PHYML v.3, que nous avons recompilé afin de pouvoir utiliser la bibliothèque MPI[15] (*Message passing interface*) et tirer profit des processeurs multi-cœurs. Le modèle ayant obtenu le meilleur AIC_c a été indiqué à PHYML *via* l'option *custom*. Par contre, les paramètres « fréquences des bases à l'équilibre », « taux de transitions/transversions », « taux de sites invariants » et « forme de la distribution Γ » estimés par JMODELTEST sont recalculés par PHYML. En effet, la topologie obtenue *via* PHYML peut différer sensiblement de celle proposée par JMODELTEST, c'est pourquoi il est nécessaire de ré-estimer ces paramètres (Stéphane Guindon, communication personnelle).

Toutes les analyses que nous avons réalisées ont nécessité l'utilisation d'une distribution Γ découpée en 4 classes centrées sur la moyenne (voir la note 11 page 80 au sujet du nombre de classes). La recherche de la meilleur topologie est faite en utilisant comme point de départ un arbre calculé par BIONJ (Gascuel, 1997). À partir de ce point, la recherche du maximum de vraisemblance a été réalisée en utilisant deux heuristiques (*nearest neighbor interchange* et *subtree pruning and regrafting*) et en gardant la meilleure des deux solutions à chaque itération, jusqu'à la stabilisation complète des paramètres (Guindon, 2003). Cette analyse a été répétée 100 fois à partir de points aléatoires afin de couvrir au mieux l'espace des topologies possibles et de ne pas passer à côté d'une solution alternative.

Comme pour les analyses non-paramétriques, nous avons utilisé la méthode du *bootstrap* (1 000 réplicats) pour évaluer la robustesse de cette reconstruction phylogénétique, mais avec cette fois un gain de temps important dû à la parallélisation[16]. En effet, la bibliothèque MPI permet de répartir les réplicats sur les processeurs présents. En théorie, n processeurs permettent de réaliser n bootstraps simultanés. En passant d'un à deux cœurs, nous avons pu observer une diminution du temps de calcul d'environ 40 %, proche du gain de 50 % attendu. Avec la généralisation des architectures multi-cœurs (quadricœurs actuellement et hexacœurs en 2010), cette stratégie devrait permettre des gains de temps très importants ou l'utilisation de modèles d'évolution plus complexes.

Malgré cette parallélisation, la méthode du *bootstrap* reste lourde et ne sera pas raisonnablement applicable aux grandes reconstructions phylogénétiques concernant plusieurs centaines ou plusieurs milliers de taxons (voir par exemple Dalevi *et al.*, 2007). Une piste intéressante est celle du calcul des valeurs d'*approximate Likelihood-Ratio Test* ou aLRT (Anisimova & Gascuel, 2006). Ces valeurs sont des estimations

15. `http://www.mcs.anl.gov/research/projects/mpich2/index.php`

16. PHYML étant un logiciel libre, il est possible de le modifier pour étendre ses possibilités, à l'inverse de PAUP. Je pense personnellement que le logiciel propriétaire n'a pas sa place en science, puisque la méthodologie sous-jacente est occultée et de fait ne peut être ni reproduite, ni critiquée.

permettant d'obtenir beaucoup plus rapidement — environ 5 % du temps total de l'analyse phylogénétique — la probabilité qu'un clade soit vrai, pour une matrice et un modèle d'évolution donnés. Les valeurs aLRT ont été générées *via* PHYML pour toutes nos analyses et comparées aux valeurs de *bootstrap*. Pour les nœuds fortement soutenus (*bootstrap* > 90 %), l'accord entre les deux techniques est bon. Par contre, nos observations montrent un comportement plus erratique de l'aLRT pour les nœuds moins bien soutenus. Nous n'avons donc pas intégré cet outil à nos résultats, mais les progrès dans ce domaine devront être suivis attentivement.

5.1.3.3 Probabilités bayésiennes

L'inférence bayésienne[17] est la démarche logique permettant de calculer ou réviser la probabilité d'une hypothèse *a priori*, à mesure que les observations sont prises en compte. Cette démarche trouve aujourd'hui de nombreuses applications pratiques, dont par exemple les filtres bayésiens utilisés pour le filtrage du *spam* sur Internet. Appliquée à la phylogénie, la méthode bayésienne est une alternative intéressante à la méthode du maximum de vraisemblance grâce notamment à des algorithmes efficaces (Larget & Simon, 1999 ; Huelsenbeck & Ronquist, 2001) et des probabilités *a posteriori* permettant de juger du soutien statistique d'un clade sans recourir à des techniques de *bootstrap*.

En parallèle aux reconstructions phylogénétiques basées sur le maximum de parcimonie et le maximum de vraisemblance, nous avons utilisé les logiciels MRBAYES v.3.5 (Huelsenbeck & Ronquist, 2001 ; Ronquist & Huelsenbeck, 2003) et BEAST v.1.4.8 (A. J. Drummond & Rambaut, 2007). Dans le cas de MRBAYES, la sélection du meilleur modèle a été réalisée par MRMODELTEST (Nylander, 2004), suivant la procédure décrite page 79. Le logiciel BEAST ne propose par défaut que les modèles Hasegawa-Kishino-Yano (HKY85) et *General Time Reversible* (GTR). Nous avons donc dupliqué chaque analyse afin de vérifier l'absence de changements topologiques liés au choix du modèle. Les analyses ont été réalisées en incluant la fonction d'hétérogénéité Γ (4 classes) et les probabilités *a priori* sont celles des réglages par défaut. Suivant les recommandations des auteurs, nous avons utilisé le paramètre *relaxed clock uncorrelated lognormal* pour autoriser une variation des taux d'évolution d'une branche à l'autre (par opposition à l'hypothèse d'« horloge moléculaire stricte » qui implique des taux d'évolution uniformes).

Nombre de générations Le choix de la longueur de la chaîne[18], c'est-à-dire du nombre de générations, et de la fréquence d'échantillonnage sont des paramètres importants. A. J. Drummond & Rambaut (2007) préconisent la relation suivante $f(taxons) = 3\,000 \times \text{taxons}^2$, soit 30×10^6 générations pour 100 taxons, et 120×10^6

17. Thomas Bayes (env. 1702, Londres – 17 avril 1761) est un pasteur de l'Église presbytérienne britannique. Ses travaux mathématiques ont été résumés dans son *Essay Towards Solving a Problem in the Doctrine of Chances* publié en 1763 à titre posthume dans les comptes-rendus de l'Académie royale de Londres (voir `http://www.stat.ucla.edu/history/essay.pdf` pour une version numérisée).

18. Le logiciel lance en réalité deux analyses en parallèle, de quatre chaînes chacunes.

générations pour 200 taxons. À chaque génération, les paramètres — topologie de l'arbre, longueurs de branches et taux de variations — sont modifiés pour converger vers la solution optimale. La vitesse de cette convergence est fortement dépendante de la taille du jeu de données et de la force du signal phylogénétique contenu dans la matrice. Pour un jeu de 100 taxons, la convergence peut se faire en beaucoup moins de 30 millions de générations mais également en beaucoup plus (et éventuellement jamais).

On le voit, une des principales difficultés de la méthode bayésienne est de décider s'il y a convergence ou non des différentes chaînes sur une seule et unique solution, d'autant que cette étape critique est souvent postérieure à l'analyse et basée le plus souvent sur une visualisation de la dispersion des chaînes à chaque échantillonnage. Il manque donc un outil basé sur des tests statistiques rigoureux et permettant de décider en temps réel de l'arrêt ou de la poursuite de l'analyse bayésienne[19]. Ce manque d'outils nous a conduit à sur-dimensionner nos analyses en réalisant systématiquement 10 millions de générations, avec échantillonnage toutes les 1 000 générations et, partant de l'hypothèse d'une convergence achevée en un million de générations, rejet des 1 000 premiers échantillons. Chaque analyse a été réitérée cinq fois pour minimiser les risques de convergence vers un optimum local et s'assurer de la répétabilité de l'expérience.

Probabilités *a posteriori* Si l'interprétation du test du *bootstrap* fait encore l'objet de discussions, les probabilités *a posteriori* calculées à l'issue d'une reconstruction bayésienne ont une définition claire. La probabilité *a posteriori* d'un clade est la probabilité que ce clade soit vrai, pour un jeux de données, un modèle d'évolution et une probabilité *a priori* données. Cependant, les bootstraps et les probabilités *a posteriori* se comportent différemment. En effet sous l'influence de certains paramètres — probabilités *a priori*, modèle d'évolution, taille du jeux de données —, les probabilités *a posteriori* tendent à surévaluer le support des clades par rapport au *bootstrap* (Suzuki *et al.*, 2002 ; Simmons *et al.*, 2004 ; Z. Yang & Rannala, 2005 et Z. Yang, 2008 pour une proposition de solution). Le *bootstrap* étant plus conservateur, c'est lui qui a été retenu pour l'interprétation des phylogénies réalisées au cours de ce travail de thèse.

5.1.3.4 Estimation de la contrainte sélective

Deux types de mutations peuvent intervenir sur une séquence codante : les mutations synonymes ou non-synonymes. L'hypothèse sous-jacente est que la sélection naturelle s'exerce sur les mutations non-synonymes puisque celles-ci impliquent des changements dans la séquence d'acides aminés, à la différence des mutations synonymes. En observant les proportions de ces deux types de mutations, il est possible d'en déduire le type de pression de sélection — neutre, positive ou négative — s'étant exercé au cours du temps. Il existe deux grandes familles de méthodes pour estimer la contrainte sélective. En fonction des objectifs fixés, il est possible de rechercher des

19. Le logiciel AWTY (Nylander *et al.*, 2008) apporte une réponse partielle à ce problème.

changements de contrainte évolutive d'une branche à l'autre ou bien d'un codon à l'autre (voir Fig. 5.5).

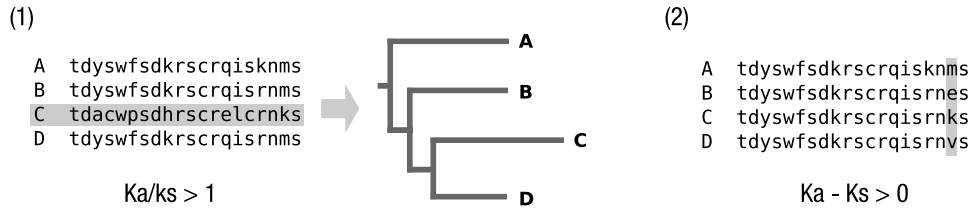

FIG. 5.5 — Estimation de la pression de sélection exercée sur des séquences codantes. Exemple de pression de sélection diversifiante sur (1) une séquence, ce qui se traduit par une branche plus longue ; (2) un codon particulier, ce qui se traduit par une variabilité importante à cette position.

Par séquence Le rapport du nombre de substitutions non-synonymes K_A et du nombre de substitutions synonymes K_S informe sur le type de sélection qui s'est exercé au cours du temps : sélection purifiante (<1), sélection neutre (=1), sélection diversifiante (>1). Ce type d'analyse est réalisé sur des paires de séquences, dans le cas d'une matrice de n séquences, chaque séquence est comparée à chacune des $n-1$ autres. Puisqu'il s'agit du calcul d'un ratio, les séquences rigoureusement identiques ont été éliminées du jeu de données à l'aide de T-COFFEE pour éviter une division par zéro.

Un grand nombre de méthodes ont été proposées pour le calcul des ratios K_A/K_S[20]. Nous nous sommes tournés vers le logiciel KAKS_CALCULATOR qui permet de tester simultanément plusieurs méthodes (Z. Zhang *et al.*, 2006, et références citées). Ce logiciel intègre notamment une méthode basée sur le maximum de vraisemblance et utilisant les outils mathématiques décrits page 79 pour classer les 14 modèles d'évolution testés. KAKS_CALCULATOR propose en sortie les résultats du modèle correspondant le mieux au données mais également les résultats moyennés de l'ensemble des modèles testés. Selon Z. Zhang *et al.*, cette dernière approche minimise les biais relatifs au choix d'un modèle, nous l'avons donc retenue pour nos analyses[21].

Par codon Plusieurs méthodes permettent de faire le calcul codon par codon pour identifier les régions sur lesquelles s'exerce une pression de sélection. Nous avons sélectionné les méthodes mises en œuvre dans la suite logicielle HYPHY (Kosakovsky Pond *et al.*, 2005) et disponibles sur le site DATAMONKEY (Kosakovsky Pond & Frost, 2005a)[22]. Cette suite logicielle se base sur des modèles d'évolution de codons pour évaluer les proportions de mutations synonymes et non-synonymes ($K_A - K_S$). En utilisant un consensus majoritaire basé sur les résultats des méthodes SLAC, FEL

20. On rencontre également l'écriture $\omega = d_N/d_S$, voir Hurst (2002) pour revue et Nekrutenko *et al.* (2002) pour une autre utilisation de ce ratio.

21. Le logiciel KAKS_CALCULATOR utilisant en entrée le format AXT, j'ai créé un convertisseur pour passer une matrice FASTA au format AXT. J'ai également créé un outil de mise en forme et de synthèse des fichiers de sortie de KAKS_CALCULATOR (code disponible sur demande).

22. `http://www.datamonkey.org/`

et REL (Kosakovsky Pond & Frost, 2005b ; Poon *et al.*, 2009), nous avons pu identifier dans nos jeux de données les codons soumis à une pression de sélection positive ou négative.

5.1.3.5 Perspectives en phylogénie

Les progrès de l'algorithmique et l'augmentation continue de la puissance des micro-ordinateurs ont permis l'utilisation de techniques de reconstruction phylogénétique de plus en plus raffinées et intégrant mieux les connaissances actuelles sur les mécanismes évolutifs. Aujourd'hui, le développement rapide de la phylogénomique pose de nouveaux défis algorithmiques, comme la constitution automatique de très grandes matrices muti-gènes et la possibilité de traiter ces matrices, mêmes incomplètes (Stamatakis & Ott, 2008).

Ces dernières années ont également vu l'apparition des premiers logiciels capables de fusionner l'étape d'alignement avec le reste des étapes d'optimisation : recherche du meilleur modèle, optimisation de la topologie et des paramètres évolutifs (Lunter *et al.*, 2005 ; Novák *et al.*, 2008 ; K. Liu *et al.*, 2009 ; Loÿtynoja & Goldman, 2009 ; Yue *et al.*, 2009). Citons par exemple le logiciel STATALIGN[23] (Novák *et al.*, 2008). Ce dernier fournit une interface graphique pour les analyses bayésiennes et propose de réaliser toutes les analyses — alignement de la matrice, codage des indels, inférence de la phylogénie et sélection du modèle d'évolution — en une seule étape. En effet, dans la démarche classique, l'étape d'alignement multiple est déconnectée de l'étape de reconstruction phylogénétique, c'est-à-dire que les choix faits à cette étape ne sont pas évalués au regard de leur impact sur la topologie finale de l'arbre. L'approche fusionnée fait disparaître ce cloisonnement artificiel et améliore la vitesse et la qualité des inférences. Selon Novák *et al.* (2008), plusieurs artéfacts sont ainsi évités, dont la très forte influence sur l'analyse phylogénétique de l'arbre-guide utilisé par le logiciel d'alignement multiple (Nelesen *et al.*, 2008).

En raison de leur nouveauté, nous n'avons testé ces méthodes que superficiellement, en comparant les résultats avec ceux obtenus *via* les techniques décrites dans ce chapitre. Cependant, il m'a semblé important de signaler ici l'avancée importante, tant sur le plan pratique que conceptuel, que représente la fusion de toutes les étapes du processus de reconstruction phylogénétique.

5.1.4 Visualisation et analyse des arbres phylogénétiques

L'arbre phylogénétique est un outil puissant, permettant de synthétiser une très grande quantité d'information en une seule image. Cependant son interprétation peut être difficile, c'est pourquoi la Fig. 5.6 page suivante propose un rappel de certaines notions et du vocabulaire associé à la phylogénie. Pour visualiser et mettre en forme les arbres générés par nos analyses, nous avons utilisé les logiciels FIGTREE[24] et TREE-

23. `http://phylogeny-cafe.elte.hu/StatAlign/`
24. `http://tree.bio.ed.ac.uk/software/figtree/`

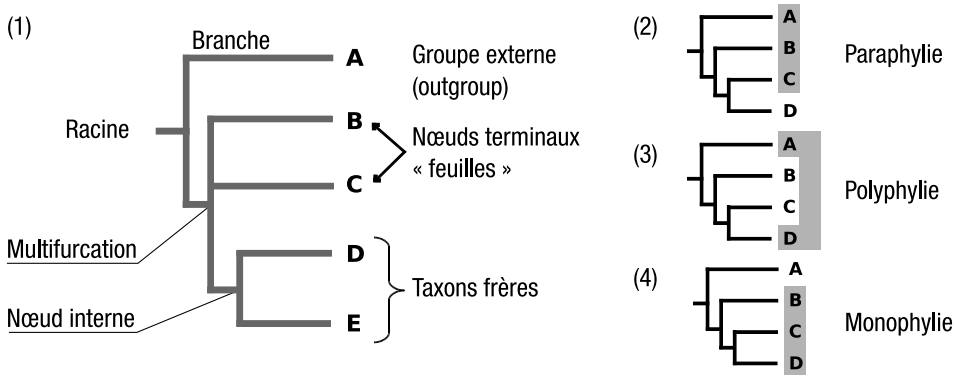

FIG. 5.6 — Illustration du vocabulaire phylogénétique (adapté de Gregory, 2008). (1) Un arbre phylogénétique est polarisé ; en remontant des feuilles vers la racine, on remonte le temps. Les nœuds terminaux sont connectés par des branches qui se rejoignent au niveau des nœuds internes. Ces nœuds internes représentent les évènements de spéciation déterminés à partir du jeu de données. L'ensemble des feuilles auxquelles on aboutit en partant d'un nœud interne est appelé « clade ». Dans notre exemple, le taxon *D* est taxon-frère de *E*. De même, le taxon *C* est frère du clade $D + E$. Le groupe externe (*outgroup*) est nécessaire pour pouvoir enraciner/polariser l'arbre. Seuls les clades monophylétiques (4) incluant tous les descendants d'un nœud sont valables phylogénétiquement. Un clade paraphylétique (3) désigne un clade incluant une espèce ancestrale et une partie seulement de ses descendants (par exemple, le groupe des « reptiles » inclut les dinosaures mais pas les oiseaux, leurs descendants). Un clade polyphylétique est un clade défini par une ressemblance qui n'a pas été héritée d'un ancêtre commun, mais qui est le fruit d'une convergence évolutive (les « ongulés » par exemple).

GRAPH[25] (Müller & Müller, 2004) qui présentent l'avantage d'être simples d'utilisation, multi-plateformes et de pouvoir exporter les arbres dans de nombreux formats dont le format vectoriel SVG (format standardisé utilisé pour générér la quasi-totalité des figures présentées dans ce manuscrit).

5.2 Annotation de BAC et génomique comparative

Les premiers travaux génomiques visant le genre *Lupinus* ont débuté récemment avec la publication de cartes génétiques (Boersma *et al.*, 2005 ; Nelson *et al.*, 2006 ; Phan *et al.*, 2007 ; Nelson *et al.*, 2008 ; Chudy *et al.*, 2008) et l'annonce de la création d'une banque BAC (*Bacterial artificial chromosome*) par le laboratoire de Bogdan Wolko (Kasprzak *et al.*, 2006). Cette banque BAC, construite selon le protocole détaillé dans Shi *et al.* (2009), couvre la quasi-totalité du génome de *Lupinus angustifolius* L. cv. *sonet*. En collaboration avec leur équipe, nous avons selectionné des BAC contenant le gène *SymRK*. L'objectif pour nous est triple : 1) obtenir la séquence complète de *SymRK* et de son entourage, 2) obtenir une première séquence génomique de référence pour l'étude des éléments transposables dans les génomes de lupins et 3) comparer cette séquence aux régions homologues chez les fabacées modèles.

25. `http://treegraph.bioinfweb.info/`

5.2.1 Criblage de la banque

La banque BAC produite par Kasprzak *et al.* (2006) correspond à six fois la taille du génome haploïde de *Lupinus angustifolius*. La banque couvre-t-elle pour autant la totalité du génome ? D'après la formule de Clarke & Carbon (1976), la probabilité P qu'un gène pris au hasard soit présent dans une banque de N clones est :

$$P = 1 - (1 - f)^N \qquad \text{avec} \qquad f = \frac{\text{taille moyenne des inserts}}{\text{taille du génome haploïde}}$$

Considérant une taille de génome haploïde de 924 Mb (Naganowska *et al.*, 2003) et une taille moyenne des inserts de 100 kb, la probabilité qu'une séquence du génome de *L. angustifolius* prise au hasard soit présente dans l'un des 55 296 clones de la banque est de 99,7 %[26].

Pour identifier les BAC porteurs du gène *SymRK*, nous avons fourni à l'équipe de Bogdan Wolko un amplifiat de la partie 5′ du gène (amorces F3-R5). Cet amplifiat a été utilisé comme une sonde marquée, selon la méthode décrite par Książkiewicz *et al.* (2008). Une PCR d'amplification du *SymRK* a été réalisée sur les six BAC identifiés. Le séquençage de ces amplifiats a confirmé qu'il s'agissait bien du gène recherché. Une digestion enzymatique a permis de vérifier que les six BAC recouvrent bien la même région génomique. Nous avons ensuite sélectionné le plus grand d'entre eux pour un séquençage selon la méthode Sanger *et al.* (1977). Le séquençage et l'assemblage de la séquence ont été réalisés par la société Agowa (Berlin, Allemagne).

5.2.2 Processus d'annotation

L'annotation de la séquence obtenue a été réalisée en recherchant trois catégories de séquences. Les outils étant nombreux, nous nous sommes focalisés sur des logiciels d'usage courant. Pour toutes les manipulations intermédiaires, conversions de formats, découpages, filtrages, condensation et préparations de séquences, des scripts ont été créés (disponibles sur demande).

5.2.2.1 Détection et annotation des régions codantes

Chez les eucaryotes, les gènes codant pour une protéine sont composés d'un site d'initiation de la transcription, d'un codon start et d'un codon stop marquant le début et la fin de la traduction et d'une séquence terminatrice qui marque la fin de la transcription. Le codon start indique le début du ou des exons. S'ils sont plusieurs, les exons sont séparés par des introns, les limites exon-intron étant indiquées par la présence de sites donneurs et accepteurs (*splice sites*). Chez les angiospermes, l'intron commence en général par une séquence GT et se termine par une séquence AG. À cette structure générale s'ajoute de nombreux cas particuliers, des changements de brin, des introns

26. Statistiquement, pour que la banque porte le gène qui nous intéresse avec une probabilité de 99 %, la taille de la banque doit faire au moins 4,6 la taille du génome. À titre de comparaison, une banque couvrant 99 % du génome du blé (16 Gb) nécessite au moins 736 000 clones ; et plus de 1 105 000 si l'on souhaite une couverture de 99,9 %.

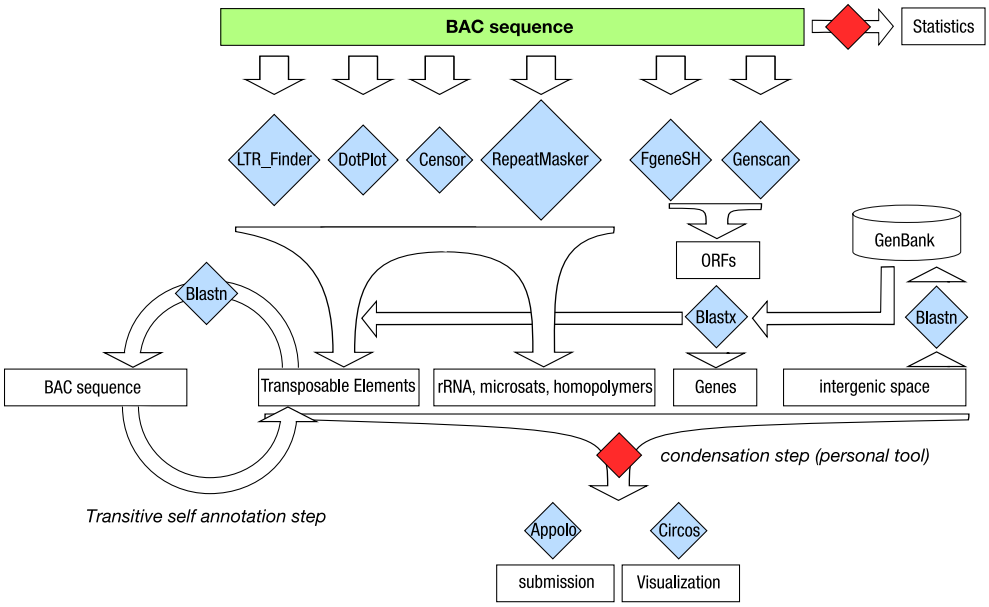

FIG. 5.7 — Processus d'annotation de BAC. Trois catégories de séquences non-exclusives ont été recherchées : les séquences codantes, les séquences répétées et les séquences conservées.

très longs et des variations dans les séquences des sites de fixation, rendant difficile l'identification exhaustive des régions codantes.

Parmi les différents logiciels testés — GENSCAN[27] (Burge & Karlin, 1997, 1998), FGENESH[28] (Salamov & Solovyev, 2000), GENEMARK HMM, GLIMMERM et NET PLANT GENE —, c'est le logiciel FGENESH qui s'est montré le plus sensible (la sélectivité étant moins importante pour nous à ce stade). FGENESH se base sur les modèles de Markov cachés (*Hidden Markov Models* ou HMM) pour identifier les régions codantes. Nous avons choisi d'utiliser un HMM entraîné sur *Medicago truncatula*, le génome-modèle le plus proche du genre *Lupinus*.

Afin d'annoter les gènes prédits par FGENESH, les séquences des régions codantes (ARNm) ont été extraites et soumises à deux analyses. Une première analyse au niveau nucléotidique (BLASTN) pour identifier des séquences hautement similaires — recherche sur les banques *nr/nt* et *dbest*, limitée aux fabacées, *e-value* fixée à 10^{-8} — et une deuxième analyse au niveau protéique, basée sur BLAST2GO[29] (Conesa *et al.*, 2005) et BLASTX pour détecter des identités plus faibles (paramètres par défaut). BLAST2GO automatise l'interrogation des tables d'annotation du NCBI, des banques de protéines (*uniprot*, *swissprot*) et de *Gene Ontology* pour annoter les ARNm prédits.

27. `http://genes.mit.edu/GENSCAN.html`
28. `http://linux1.softberry.com/berry.phtml`
29. `http://www.blast2go.de/`

5.2.2.2 Détection de séquences répétées

Chez les angiospermes, les séquences répétées peuvent représenter une grande partie du génome (cf. Chap. 1). Ces répétitions peuvent être des copies multiples de gènes (familles multigéniques) ou d'ARN (ARNt et ARNr), des séquences satellites (midi, mini ou micro-satellites), ou des éléments transposables (cf. Chap. 2). Ces séquences représentent un ensemble extrêmement diversifié et sont particulièrement difficiles à identifier de façon exhaustive.

Plusieurs approches peuvent être utilisées pour la recherche de séquences répétées mais la méthode la plus courante est sans conteste celle du dictionnaire. La séquence inconnue est comparée à une banque de séquences représentatives de différentes familles d'éléments répétés. Très efficace, cette méthode ne permet toutefois que d'identifier ce qui est déjà connu par ailleurs. Ce manque de flexibilité peut être contrebalancé par la construction de « profils » correspondant à des familles de séquences, ou de modèles décrivant de façon abstraite ce qui caractérise les séquences ciblées (structures secondaires, motifs génériques). Nous avons utilisé trois outils basés sur ces principes : REPEATMASKER[30], CENSOR[31] (Kohany *et al.*, 2006) et LTR FINDER[32]. Pour REPEATMASKER (versions 3.2.6 et 3.2.7), afin d'obtenir une sensibilité maximale, nous avons utilisé le moteur *cross match* et l'option *slow*, en spécifiant *Medicago truncatula* comme origine de la séquence. Nous avons également utilisé l'outil *protein based repeatmasker* pour les identités plus lointaines. CENSOR a été utilisé en ciblant les répétitions identifiées chez les Viridiplantae, en utilisant également le niveau protéique. De plus, disposant d'une banque de 380 séquences de *reverse transcriptase* de lupins (voir Chap. 7), nous avons confronté cette banque à la séquence du BAC via BLASTN (paramètres par défaut).

En parallèle à cette approche par dictionnaire, nous avons également utilisé une méthode biologiquement plus naïve, basée sur une définition formelle de ce qu'est une répétition :

> Une séquence répétée est une suite de n nucléotides, présente au moins deux fois dans une ou plusieurs suites de nucléotides de tailles supérieures ou égales à $n + 1$. Les répétitions peuvent être directes, inverses, inverses-compléments, palindromiques ou emboîtées. Cette définition simple peut être étendue à des répétitions inexactes, comprenant des erreurs.

Pour rechercher des séquences répondant à cette définition, nous avons commencé par comparer la séquence du BAC à elle-même, ce qui équivaut à un alignement de deux séquences identiques. Les méthodes d'alignement global donnent une réponse simple à cette question : la séquence s'aligne parfaitement sur elle-même. Les méthodes d'alignement local apportent une information supplémentaire. En découpant la séquence en blocs plus petits et en procédant par « ancrage et extension » — *seed & extend*, c'est typiquement le mode de fonctionnement de BLAST —, ils permettent d'identifier d'éventuelles occurrences multiples d'un même bloc. Pour réaliser cette

30. `http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker`
31. `http://www.girinst.org/censor/index.php`
32. `http://tlife.fudan.edu.cn/ltr_finder/`

opération, nous avons utilisé l'outil DOTPLOT[33] (Church & Helfman, 1993 ; Ovcharenko *et al.*, 2004).

Dans la même optique, les régions du BAC n'ayant pas reçu d'annotation dans les étapes précédentes ont été extraites et traitées par BLASTN (recherche de séquences hautement similaires sur la banque *nr/nt* limitée aux fabacées, *e-value* fixée à 10^{-8}). L'objectif étant de mettre en évidence des régions de l'espace intergénique conservées d'une espèce à l'autre (*conserved non-coding sequences* ou CNCS) et donc susceptibles de jouer un rôle important, dont celui de régulateur de l'expression des gènes (Peterson *et al.*, 2009).

La totalité de ces informations ont été combinées et intégrées dans un tableau d'annotations pour les étapes de visualisation (CIRCOS[34]) et de soumission à GENBANK (TBL2ASN[35]).

5.2.3 Recherche de régions homologues

Afin d'affiner l'étude de cette première séquence génomique de lupin, nous l'avons comparé aux régions homologues disponibles pour d'autres organismes. À la mi-2008, les projets de séquençage de génomes avaient été annoncés pour les légumineuses suivantes :

- *Medicago truncatula*
- *Lotus japonicus*
- *Glycine max*
- *Phaseolus coccineus*
- *Phaseolus vulgaris*
- *Vigna radiata*
- *Medicago sativa*
- *Cajanus cajan*
- *Melilotus albus*
- *Trifolium pratense*
- *Cicer arietinum*
- *Lens culinaris*
- *Pisum sativum*

ou pour des espèces plus éloignées :

- *Populus trichocarpa*
- *Carica papaya*

Des banques de séquences ou des assemblages partiels n'étaient disponibles ou ne sont devenues disponibles qu'en 2009 pour *Medicago truncatula*[36], *Lotus japonicus* (Sato *et al.*, 2008), *Glycine max*[37] et *Populus trichocarpa*[38] (Tuskan *et al.*, 2006 ; Jansson & Douglas, 2007). Ces différentes banques ont été interrogées en utilisant les outils de BLAST mis en place sur les sites web dédiés. L'identification de régions homologues a été faite en recherchant dans ces données génomiques les gènes présents sur notre BAC. Les régions homologues identifiées ont ensuite été extraites, annotées et comparées selon la méthode décrite plus haut.

33. http://zpicture.dcode.org/
34. http://mkweb.bcgsc.ca/circos/?
35. http://www.ncbi.nlm.nih.gov/Genbank/tbl2asn2.html
36. http://www.medicago.org/genome/
37. http://www.phytozome.net/soybean
38. http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html

Troisième partie

Résultats

6

Phylogénie moléculaire du genre *Lupinus*

« . . . that the characters which naturalists consider as showing true affinity between any two or more species, are those which have been inherited from a common parent, and, in so far, ***all true classification is genealogical****; that community of descent is the hidden bond which naturalists have been unconsciously seeking, and not some unknown plan of creation, or the enunciation of general propositions, and the mere putting together and separating objects more or less alike. »*

Charles Darwin (1859)

BIEN que des progrès considérables aient été réalisés au cours de cette dernière décennie, la phylogénie du genre *Lupinus* pose encore un certain nombre de difficultés (voir section 3.3 page 47). Notre objectif étant d'améliorer la compréhension de l'histoire des lupins, nous avons complété les jeux de données existant pour les ITS (*Internal transcribed sequence*) et les ETS (*External transcribed sequence*), deux régions transcrites du gène nucléaire répété codant les ARN ribosomiques (voir page 62). À ces séquences déjà utilisées pour la reconstruction phylogénétique, nous avons ajouté celles du gène nucléaire *SymRK* (*Symbiosis receptor-kinase*).

Grâce aux amorces universelles conçues par White *et al.* (1990), les ITS ont été largement utilisés et ont contribué à établir les relations au sein de nombreux groupes de plantes (Baldwin *et al.*, 1995 ; Soltis & Soltis, 1998). Les ETS se sont montrés plus informatifs au niveau générique et inter-spécifique chez plusieurs groupes, y compris chez les fabacées (Baldwin & Markos, 1998 ; Bena *et al.*, 1998 ; Chandler *et al.*, 2001 ; Urbatsch *et al.*, 2003 ; Suárez-Santiago *et al.*, 2007). Cependant, comme d'autres familles de gènes en multicopies, les ARN ribosomiques sont sujets à de l'évolution concertée, un processus qui tend à homogénéiser les différentes copies (*via* notamment des recombinaisons non-homologues)[1], et peut ainsi conduire à des reconstructions phylogénétiques biaisées ou incorrectes (Álvarez & Wendel, 2003 ; Bailey *et al.*, 2003 ; Kovařík *et al.*, 2005).

1. L'évolution des régions codant pour les ARN ribosomiques peut être très complexe dans le cas d'une allopolyploïdie (Dadejová *et al.*, 2007).

Toutefois, l'accumulation de connaissances sur la dynamique évolutive des gènes répétés ARNr, permet actuellement une meilleure exploitation phylogénétique des données fournies par ces régions, y compris pour mettre en évidence des processus d'évolution réticulée (Soltis *et al.*, 2008). De fait, les espaceurs de l'ARNr, en particulier les ITS restent aujourd'hui les séquences les plus utilisées pour la reconstruction des liens de parenté entre taxons proches (espèces, genres).

En complément des régions ITS et ETS, Nous avons également utilisé le gène nucléaire *SymRK* (voir page 65), supposé en faible nombre de copies et jouant un rôle important chez les légumineuses. Le gène *SymRK* code pour une protéine transmembranaire permettant la symbiose entre la plante et les bactéries fixatrices d'azote. Cette protéine est divisée en trois régions : une région réceptrice extracellulaire, un domaine intramembranaire et un domaine intracellulaire de type kinase.

Dans une démarche de *total évidence* (Kluge, 1998 ; Lecointre & Deleporte, 2005 ; Cotton & Wilkinson, 2009), les matrices ITS, ETS et *SymRK* ont été combinées afin de maximiser le signal phylogénétique[2].

6.1 Variabilité des séquences utilisées

Dans le tableau 6–1 sont regroupées quelques données statistiques concernant les matrices ITS, ETS et *SymRK* analysées dans la suite du chapitre. Les *outgroups* ont été exclus afin que les données présentées soient représentatives du genre *Lupinus*.

TAB. 6–1 — Résumé des données statistiques sur les matrices ITS, ETS et *SymRK*. La distance utilisée est la distance *p* non-corrigée (une base identique compte pour zéro et une base différente compte pour 1).

Gène	**Taxons**	**Longueur**	**Caractères informatifs**	**Distance entre les séquences (%)**			
				min	**max**	**moy**	**médiane**
ITS	48	461–466	62	0,00	6,90	3,37	3,66
ETS	46	330–345	117	0,00	22,01	10,66	11,28
SymRK	40	1 410–1 479	222	0,00	7,99	4,70	5,06

L'ITS présente environ 13 caractères informatifs pour 100 nucléotides, tandis que l'ETS, largement plus riche en signal phylogénétique, présente environ 34 caractères informatifs pour 100 nucléotides. Le *SymRK* présente un niveau de signal comparable à celui de l'ITS (15 caractères informatifs pour 100 nucléotides). Toutefois, la portion de *SymRK* étudiée étant plus longue, celle-ci fournit au final près de deux fois plus de caractères informatifs que les ETS (222 contre 117).

2. Certaines séquences ont été tirées de la banque de donnée *GenBank* (Benson *et al.*, 1994–2009).

6.2 Phylogénies des espaceurs transcrits de l'ARNr, régions ITS et ETS

6.2.1 Phylogénie des ITS

Au cours de cette étude, les méthodes de maximum de parcimonie et de maximum de vraisemblance ont été utilisées simultanément afin de vérifier leur accord sur les résultats et de détecter d'éventuels artefacts liés à une méthode particulière. Le maximum de vraisemblance produisant de manière générale des résultats plus précis (Guindon & Gascuel, 2003, et références citées), nous présenterons ici uniquement les phylogénies construites à l'aide des méthodes paramétriques.

La matrice ITS étudiée contient 50 séquences de lupins, se rapportant à 40 espèces et sous-espèces, et 4 séquences de taxons externes (*Chamaecytisus mollis*, *Genista tinctoria*, *Ulex parviflorus* ssp. *parviflorus* et *Ulex australis* ssp. *welwichianus*). L'ajout de ces quatre taxons porte le nombre de caractères informatifs de 62 à 80 (pour une matrice de 518 sites nucléotidiques). La matrice ITS a été soumise à JMODELTEST pour identifier le modèle d'évolution moléculaire décrivant au mieux les données (voir Tab. 6–2).

TAB. 6–2 — Sélection des six modèles ayant un $\Delta < 10$ pour la matrice ITS. Le modèle TIM3+Γ domine largement l'analyse avec un poids de plus de 87 %.

Modèle	ℓ	K	AIC_c	Δ	poids	poids cumulé
TIM3+Γ	2 380,3920	113	5 035,9520	0,0000	0,8772	0,8772
TIM3+I + Γ	2 381,3930	114	5 040,9199	4,9679	0,0732	0,9504
GTR+Γ	2 381,8414	115	5 044,7939	8,8419	0,0105	0,9609
GTR+I + Γ	2 380,4327	116	5 044,9653	9,0133	0,0097	0,9706
TrN+I + Γ	2 385,1579	113	5 045,4838	9,5318	0,0075	0,9781
TrN+Γ	2 386,6495	112	5 045,5123	9,5603	0,0074	0,9854

L'intervalle de confiance à 95 % n'inclut que deux variantes du même modèle, le poids de la première variante étant très largement dominant : TIM3+Γ ($\ell = 2\,380,3920$ et $w = 0,8772$)[3] et TIM3+I + Γ ($\ell = 2\,381,3930$ et $w = 0,0732$). Le paramètre $\alpha(\Gamma) = 0,4348$ indiquant une forte hétérogénéité des taux de variation alors que le paramètre « positions invariantes » n'influe que très peu le modèle ($p_{inv}(\mathrm{I}) = 0,0010$). La topologie moyenne pondérée, basée sur la totalité des modèles, indique que tous les clades reçoivent un support de 1. Ils sont donc stables et faiblement influencés par le choix du modèle. Le résultat de l'analyse phylogénétique et les valeurs de *bootstrap* sont présentés dans la Fig. 6.1 page suivante.

3. TIM3 pour *transitional model* (Posada, 2008). Il s'agit d'un modèle à $k + 6$ paramètres : fréquences des bases libres, taux de transversions partiellement fixés A ⇔ C = G ⇔ C et A ⇔ T = G ⇔ T, taux de transitions A ⇔ G et C ⇔ T libres.

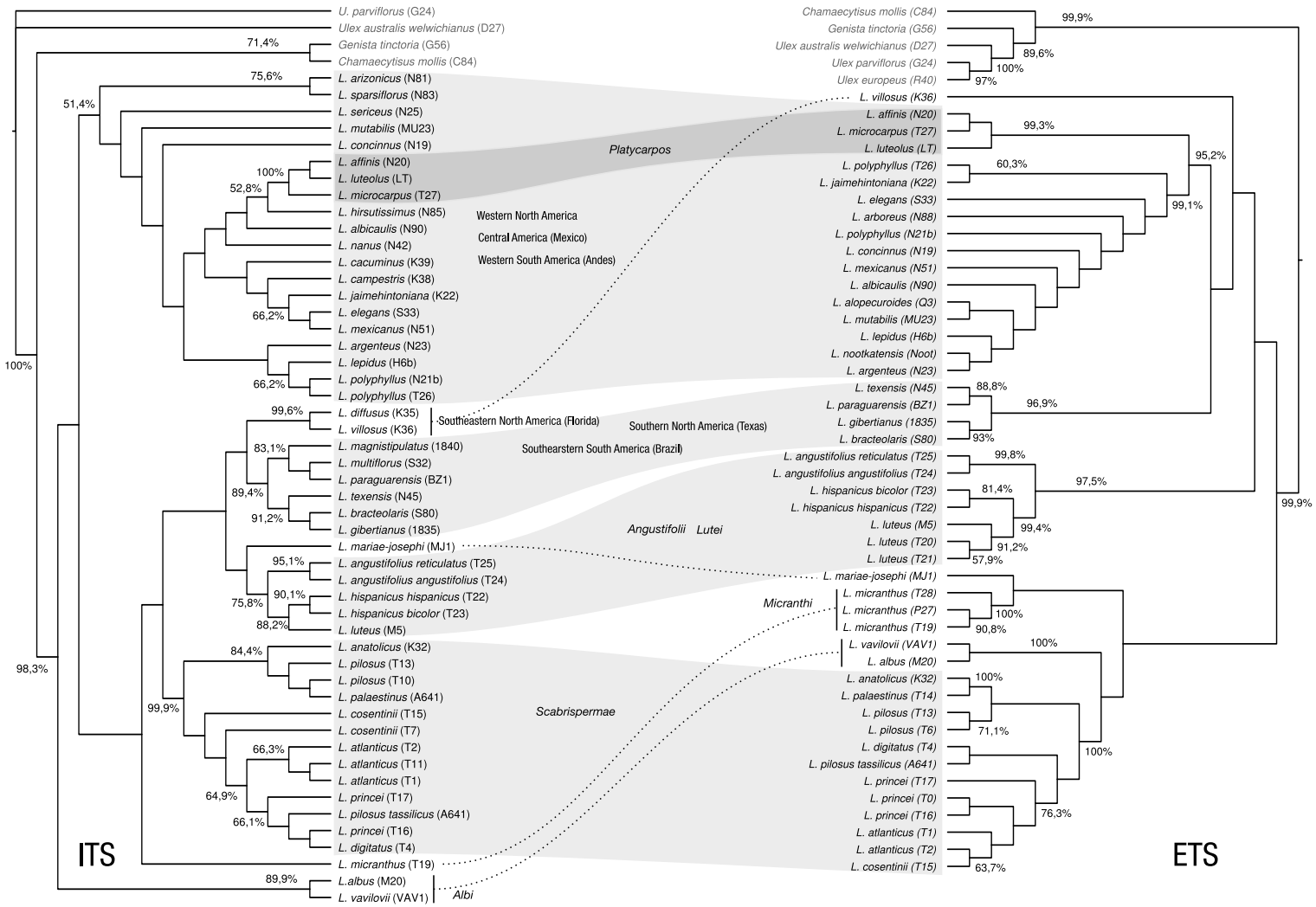

FIG. 6.1 — Phylogénies comparées des espaceurs transcrits de l'ARN ribosomique, régions ITS et ETS. Les analyses ont été réalisées selon la méthode du maximum de vraisemblance et le support des clades a été évalué par la méthode du *bootstrap* (1 000 réplicats, valeurs indiquées en pourcentage).

À l'échelle infra-générique à laquelle nous nous plaçons, la région ITS n'est pas porteuse d'un fort signal phylogénétique (80 caractères informatifs sur 518). Cependant, l'ITS confirme la monophylie du genre *Lupinus* (98,3 %) et scinde les lupins en deux grands ensembles (non-soutenus et excluant *L. albus*). Le premier groupe (51,4 %) réunit les lupins de l'Ouest de l'Amérique du Nord, d'Amérique centrale et de l'Ouest de l'Amérique du Sud (voir carte page 43). Au sein de ce groupe, les espèces de la section *Platycarpos* (Amérique du Nord) — *L. affinis*, *L. luteolus* et *L. microcarpus* — forment un clade recevant un soutien de 100 %.

Les lupins du Sud-Est de l'Amérique du Nord — *L. villosus* et *L. diffusus* (99,6 %) — et les lupins unifoliolés de l'Est de l'Amérique du Sud et leurs alliés du Sud des États-Unis (89,4 %) sont réunis avec les espèces de l'Ancien Monde au sein d'un deuxième grand groupe (rapprochement non-soutenu par le *bootstrap*). Les taxons natifs de l'Ancien Monde forment deux sous-clades bien soutenus (voir cartes pages 45 et 46), le premier regroupant les lupins des sections *Angustifolii* et *Lutei* (75,8 %) et le second les lupins à graines rugueuses (*Scabrispermae*, 99,9 %).

Le statut incertain des taxons *L. pilosus tassilicus* et *L. vavilovii* est clarifié par les ITS. *Lupinus pilosus tassilicus* ne se place pas avec les autres représentants de *L. pilosus*, mais semble appartenir à un groupe *L. digitatus*–*L. princei* (66,1 %). De même, *L. vavilovii* se groupe avec *L. albus* (89,9 %), ce qui confirme leur synonymie (Gladstones, 1974 ; Przyborowski & Weeden, 2001). Par contre, les ITS ne permettent pas de clarifier les relations entre les grands groupes de lupins. La position des taxons de l'Ancien Monde *L. albus*, *L. micranthus* et *L. mariae-josephi* reste également irrésolue.

6.2.2 Phylogénie des ETS

La matrice ETS étudiée contient 46 séquences de lupins, se rapportant à 40 espèces et sous-espèces, et 5 séquences de taxons externes (*Chamaecytisus mollis*, *Genista tinctoria*, *Ulex europeus*, *Ulex parviflorus* ssp. *parviflorus* et *Ulex australis* ssp. *welwichianus*). L'ajout de ces cinq taxons porte le nombre de caractères informatifs de 117 à 204, pour une matrice de 507 sites nucléotidiques. La matrice ETS a été soumise à JMODELTEST pour identifier le modèle d'évolution moléculaire décrivant au mieux les données. L'intervalle de confiance à 95 % inclut douze modèles (voir Tab. 6–3 page suivante), le premier étant TIM3ef+Γ ($\ell = 3\,070,7471$) avec un poids de 44 % ($w = 0,4466$)[4].

L'évaluation de l'importance des paramètres souligne le rôle joué par les différentes transversions et transitions (à l'exception de la transversion A $\Leftrightarrow$ T) et par $\alpha(\Gamma)$ dans la description de l'évolution des séquences étudiées. Environ 30 % des sites sont invariants quand l'hétérogénéité des taux n'est pas prise en compte mais cette valeur tombe à zéro quand on autorise de l'hétérogénéité. La topologie moyenne pondérée, basée sur la totalité des modèles, indique que tous les clades reçoivent un support de 1. Ils sont donc stables et faiblement influencés par le choix du modèle. Le résultat de l'analyse phylogénétique et les valeurs de *bootstrap* sont présentés dans la Fig. 6.1 page ci-contre.

4. TIM3ef, pour *equal frequencies*, est une variante à $k + 3$ paramètres de TIM3 dans laquelle les fréquences des bases ACGT sont fixes (Posada, 2008).

TAB. 6–3 — Sélection de modèles pour la matrice ETS. Sur cet exemple, le modèle produisant l'arbre le plus vraisemblable ($\ell = 3\,066,1861$) n'est pas classé premier. Il est surclassé par le modèle TIM3ef+Γ, plus économe en nombre de paramètres (K).

Modèle	ℓ	K	AIC_c	Δ	poids	poids cumulé
TIM3ef+Γ	3 070,7471	104	6 439,0025	0,0000	0,4466	0,4466
TrN+Γ	3 068,1240	106	6 441,9836	2,9812	0,1006	0,5472
SYM+Γ	3 068,1496	106	6 442,0347	3,0322	0,0981	0,6452
TIM3+Γ	3 066,1861	107	6 442,2726	3,2701	0,0871	0,7323
TIM3ef+I + Γ	3 070,7474	105	6 443,0997	4,0972	0,0576	0,7898
TIM2ef+Γ	3 073,4133	104	6 444,3349	5,3324	0,0310	0,8209
TIM2+Γ	3 067,2407	107	6 444,3819	5,3794	0,0303	0,8512
TIM1+Γ	3 067,2519	107	6 444,4043	5,4018	0,0300	0,8812
TPM3+Γ	3 075,6156	103	6 444,6760	5,6736	0,0262	0,9074
TrNef+Γ	3 075,7004	103	6 444,8457	5,8432	0,0240	0,9314
TIM1ef+Γ	3 074,2848	104	6 446,0778	7,0753	0,0130	0,9444
SYM+I + Γ	3 068,1496	107	6 446,1997	7,1972	0,0122	0,9566

L'utilisation des ETS confirme les groupes définis par les ITS et améliore leurs valeurs de soutien. Cependant, comme pour les ITS, les relations entre grands groupes de lupins restent non-résolues. Les lupins du Nouveau Monde forment un seul et unique clade (non-soutenu) composé des lupins unifoliolés d'Amérique du Sud et de leurs alliés du Texas (96,9 %), et d'un groupe soutenu à 92,5 % comprenant les *Platycarpos* (99,3 %) et les lupins de l'Ouest (99,1 %). Les lupins unifoliolés du Sud-Est de l'Amérique du Nord ne sont pas placés à proximité des lupins unifoliolés d'Amérique du Sud mais se branchent à la base du clade « Nouveau Monde » (relation non-soutenue). Dans l'Ancien Monde, les lupins des sections *Angustifolii* et *Lutei* (97,5 %) se placent en groupe-frère du clade « Nouveau Monde ». La position des autres taxons de l'Ancien Monde évolue puisque *L. albus* se place en groupe-frère des *Scabrispermae* (relation non-soutenue) et que *L. mariae-josephi* se place avec *L. micranthus* (relation non-soutenue).

6.2.3 Phylogénies combinées des ITS et ETS

Les jeux de données ITS et ETS ne générant pas d'incongruences significatives, ils ont été fusionnés pour maximiser le nombre de caractères informatifs. La matrice ITS + ETS étudiée contient 39 séquences de lupins, se rapportant à 31 espèces et sous-espèces, et 4 séquences de taxons externes (*Chamaecytisus mollis*, *Genista tinctoria*, *Ulex europeus*, *Ulex parviflorus* ssp. *parviflorus* et *Ulex australis* ssp. *welwichianus*). PHYML dans sa version actuelle ne permettant pas de découper une matrice en tronçons et d'appliquer un modèle différent à chacun d'eux, nous avons donc effectué une nouvelle recherche de modèle à l'aide de JMODELTEST pour cette matrice combinée (voir Tab. 6–4 page suivante).

TAB. 6–4 — Sélection de modèles pour la matrice ITS + ETS. Sur les 88 modèles, 18 ont un $\Delta < 10$ et les 10 premiers représentent plus de 95 % du poids des différents modèles. Les quatre premiers modèles, de poids très proches, sont des variantes de TrN ou de TIM2.

Modèle	ℓ	K	AIC_c	Δ	poids	poids cumulé
TrN+I + Γ	5 028,0023	91	10 260,6315	0,0000	0,1777	0,1777
TIM2+I + Γ	5 026,8124	92	10 260,7803	0,1488	0,1649	0,3426
TrN+Γ	5 029,3405	90	10 260,7863	0,1547	0,1644	0,5070
TIM2+Γ	5 028,2066	91	10 261,0403	0,4088	0,1448	0,6518
TIM1+I + Γ	5 027,4927	92	10 262,1410	1,5095	0,0835	0,7353
TIM1+Γ	5 028,9993	91	10 262,6255	1,9940	0,0656	0,8009
TIM3+I + Γ	5 027,9671	92	10 263,0898	2,4583	0,0520	0,8528
TIM3+Γ	5 029,3395	91	10 263,3060	2,6745	0,0466	0,8995
GTR+Γ	5 026,9210	93	10 263,5331	2,9015	0,0416	0,9411
GTR+I + Γ	5 026,3498	94	10 264,9329	4,3014	0,0207	0,9618

L'intervalle de confiance à 95 % inclut dix modèles, dont les quatre premiers, de poids très proches ($w_1 = 0,1777$, $w_2 = 0,1649$, $w_3 = 0,1644$ et $w_4 = 0,1448$), sont soit des variantes de TrN[5] (K. Tamura & Nei, 1993) soit des variantes de TIM2 (Posada, 2008). L'évaluation de l'importance des paramètres souligne le rôle joué par les transversions A ⇔ G et G ⇔ T, et la transition C ⇔ T dans la description de l'évolution des séquences étudiées. La topologie moyenne pondérée, basée sur la totalité des modèles, indique que la plupart des clades reçoivent un support de 1 (figure non-présentée). Cependant, les nœuds rattachant *L. albus*, *L. micranthus* et les lupins à graines rugueuses reçoivent un faible support (0,4), ils sont donc dépendants du choix du modèle utilisé pour optimiser la topologie de l'arbre.

La fusion des deux jeux de données n'apporte pas d'amélioration dans la résolution des relations entre les grands groupes de lupins. On notera cependant que cette analyse regroupe les lupins unifoliolés du Sud-Est de l'Amérique du Nord, les lupins unifoliolés du Sud-Est de l'Amérique du Sud et leurs alliés du Texas avec les sections *Angustifolii* et *Lutei* de l'Ancien Monde (non-soutenu). Les groupes soutenus par les analyses ITS et ETS séparées sortent renforcés de la fusion des deux jeux de données. C'est le cas par exemple des lupins de l'Ouest du Nouveau Monde (99,5 %) et des *Scabrispermae* de l'Ancien Monde (100 %).

Afin d'améliorer le cadre phylogénétique obtenu avec les espaceurs transcrits de l'ARNr, nous nous sommes orientés vers un gène nucléaire jouant un rôle clé dans les processus de symbiose entre plantes et champignons et entre plantes et bactéries : le gène *SymRK*.

5. Le modèle retenu, TrN ou TN93, est un modèle à $k + 5$ paramètres dans lequel les taux de transversions sont tous égaux, alors que les taux de transitions A ⇔ G et C ⇔ T sont libres (K. Tamura & Nei, 1993).

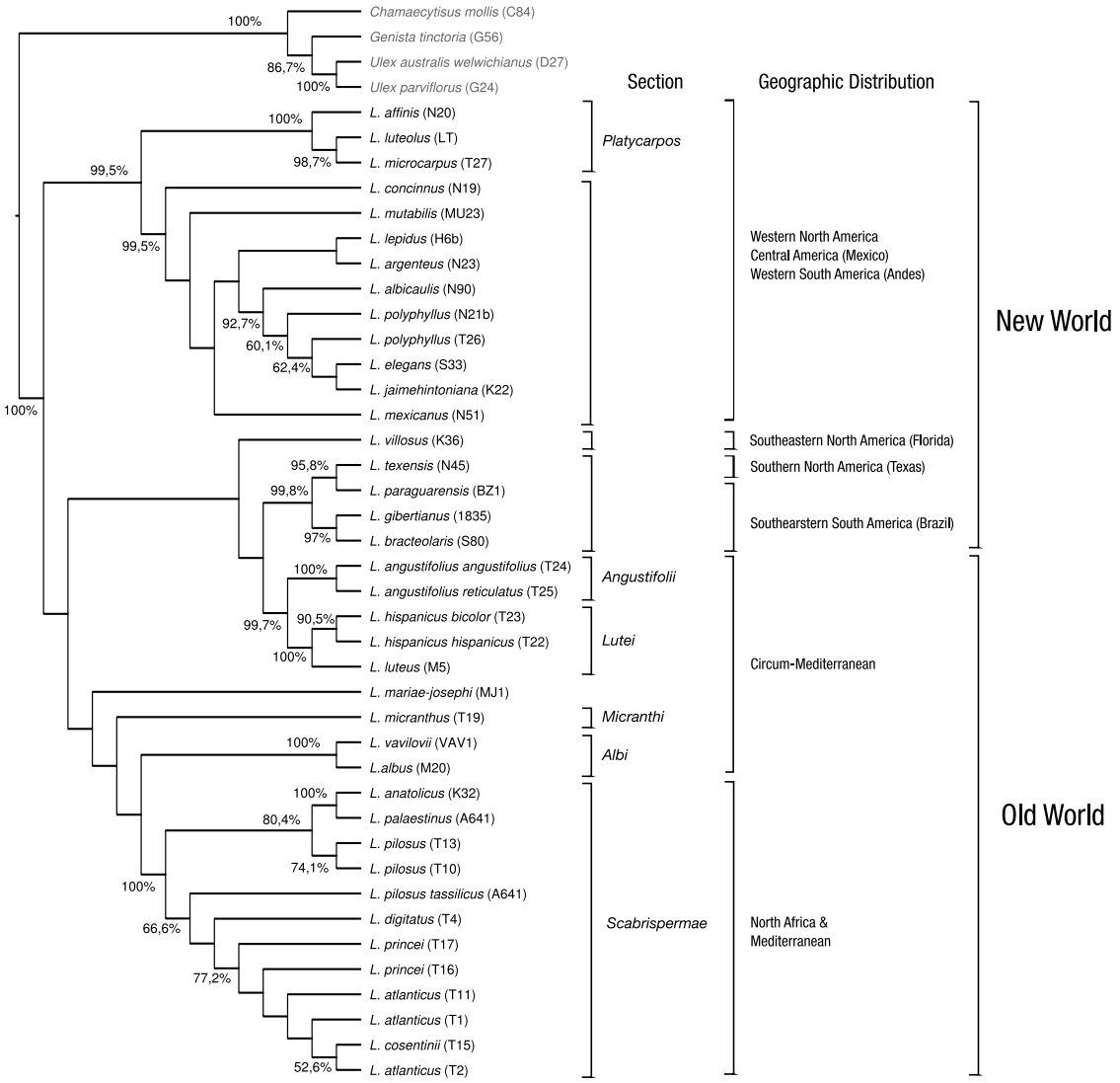

FIG. 6.2 — Phylogénie combinée des espaceurs transcrits de l'ARN ribosomique, régions ITS et ETS. Les analyses ont été réalisées selon la méthode du maximum de vraisemblance et le support des clades a été évalué par la méthode du *bootstrap* (1 000 réplicats, valeurs indiquées en pourcentage).

6.3 Phylogénie du gène *SymRK*

La majorité des plantes terrestres peuvent établir des interactions avec des champignons ou des bactéries du sol *via* leurs racines. Les plantes appartenant au groupe des Fabaceae, ou légumineuses, ont la particularité supplémentaire de pouvoir s'associer à des bactéries fixatrices d'azote atmosphérique, les *Rhizobium*. Cet apport confère aux Fabaceae un avantage important sur les autres plantes, notamment sur des sols pauvres en azote. L'établissement de cette symbiose entre Fabaceae et *Rhizobium* passe par une cascade d'interactions chimiques regroupées sous le terme « programme symbiotique ». Le gène *SymRK*, codant pour une protéine réceptrice transmembranaire, joue un rôle clé dans cette cascade. L'article présenté ci-dessous décrit l'isolement et le séquençage de la région 5′ du *SymRK* (région extra-membranaire) chez 30 lupins et 2 ajoncs. L'objectif pour nous étant de caractériser cette région, d'évaluer son intérêt pour la phylogénie du genre *Lupinus* et de détecter d'éventuels liens entre des variations de séquence de *SymRK* et des changements de symbiontes.

Les séquences de *SymRK* obtenues chez les lupins étudiés se sont révélées très proches structurellement des séquences déjà connues chez d'autres Fabaceae. De plus, l'analyse des ratios K_A / K_S indique qu'une pression de sélection purifiante s'exerce sur les régions codantes du gène *SymRK*. La reconstruction phylogénétique du *SymRK* a montré un net renforcement des groupes déjà reconnus par les études précédentes (Käss & Wink, 1997a ; Aïnouche *et al.*, 2004 ; Hughes & Eastwood, 2006 ; Ch. S. Drummond, 2008) et apporte quelques éclairages nouveaux sur les relations au sein de certains groupes. La phylogénie obtenue répartit les lupins en deux grands clades, correspondant respectivement à l'Ancien et au Nouveau Monde. Cette topologie a été confirmée par une analyse basée sur la méthode bayésienne. Cette même analyse a permis de dater, sur la base d'une divergence entre *Lupinus* et le reste des génistées il y a environ 16 millions d'années, le dernier ancêtre commun des lupins actuels à environ 6 à 7 millions d'années.

Le gène *SymRK* s'est avéré performant en tant que marqueur moléculaire pour l'analyse phylogénétique, en comparaison avec la plupart des séquences nucléaires et chloroplastiques précédemment utilisées. Par ailleurs, la région extra-cellulaire s'est révélée très conservée et soumise à une forte sélection purifiante. En effet, nous n'avons pas mis en évidence de changements adaptatifs en relation avec la diversité et la spécificité symbiotique des lupins. D'autres gènes intervenant dans le programme symbiotique, tels que les *nfr* (*node factor recognition*), pourraient s'avérer de meilleurs candidats pour aborder cette question.

Isolation, phylogeny and evolution of the *SymRK* gene in the legume genus *Lupinus* L.

Frédéric Mahé[1,a], Dragomira Markova[2,a], Marie-Thérèse Misset[1], Abdelkader Aïnouche[1]

[1]UMR CNRS 6553 Ecobio, Université de Rennes-1, Campus scientifique de Baulieu Bât 14A, F-35042 Rennes cedex, France.
[2]Institute for Genomics and Systems Biology, University of Chicago, 60637 Illinois, USA.
[a]FM and DM equally contributed to this paper.

The capacity to form symbiotic associations with fungi is widespread in land plants. Less common is Legumes' ability to accommodate nitrogen-fixing bacteria in specific root structures called nodules. The resulting increased nitrogen uptake confers to the plant an adaptive advantage, and an ecologically and economically important role of nitrogen supplier for the environment. Genetic analyses of initial steps of both symbiotic processes have shown that the gene Symbiotic Receptor-like kinase (*SymRK*) plays a key role. We isolated and sequenced a large part of the DNA sequence encoding the extracellular domain of the *SymRK* gene in 38 accessions from the genus *Lupinus* (Fabaceae; Genisteae) and 2 outgroups from the closely related genus *Ulex*. Our objective was to characterize this region and its sequence variability within the genus *Lupinus*; reconstruct the *SymRK* phylogeny and evaluate its utility for inference of the evolutionary history of lupines; and examine whether correlations can be established between the molecular pattern of divergence of the *SymRK* extracelular domain and the rhizobial diversity and specificity observed in the genus. *SymRK* appeared to be a structurally well-conserved gene under a strong purifying selection, and present in only one copy in all tested lupines. The partial 5' region used for phylogenetic analyses largely improved over previous inferences, with only a few relations left unresolved. No signature of diversifying selection was found in lupine lineages that experienced changes of symbionts, indicating that *SymRK* may not be directly responsible for rhizobial specificity in *Lupinus*.

Introduction

The ability of organisms to establish associations of mutual benefit (or symbioses) has played a fundamental role in the evolution and diversification of species. Land colonization by plants took place 400 to 500 Myr ago and was accompanied by the establishment of ubiquitous complex symbioses between fungi (Glomeromycota) and plants, especially the intracellular association (endosymbiosis) known as the arbuscular mycorrhiza (AM) (Remy et al., 1994; Schüssler et al., 2001). In this interaction, the fungus penetrates root cells and provides the plant with water and soil nutrients (particularly kalium and phosphate), which are crucial for the plant development and represent the main limiting factors for the functioning and productivity of natural and agricultural terrestrial ecosystems.

Among other less widespread mutualistic root symbioses, one of the most important events occurring in the recent evolutionary time scale (less than 100 Myr ago), was the association of plants and soil nitrogen-fixing bacteria, accompanied by the formation of specific root structures called nodules (Soltis et al., 1995). Within these endosymbiotic structures, the plant roots accommodate a protective niche and provide energy to the bacteria (mainly organic carbon produced by photosynthesis), while the nitrogen-fixing bacteria provide their host with important amounts of reduced nitrogen (ammonium), another limiting factor of plant growth

and development. Consequently, nodulating plants are independent from other nitrogen sources and themselves become nitrogen suppliers to the ecosystems. However, compared to the ubiquitous AM, the nitrogen-fixing root nodule symbioses are phylogenetically restricted to only four orders in the Eurosid I clade of eudicots in the flowering plants (Soltis et al., 1995): three orders associated with Gram-positive actinobacteria of the genus *Frankia* (Pawlowski and Sprent, 2008), Fagales, Cucurbitales and Rosales; and within the Fabales, the legume family (Fabaceae), which is notorious for its particular symbiotic association with divers Gram-negative nitrogen-fixing bacteria, collectively known as rhizobia (Pawlowski and Sprent, 2008).

With about 20,000 described species, legumes represents the third largest family of flowering plants and is characterized by a high biological and ecological diversity (Lewis et al., 2005; Lavin et al., 2005; Cronk et al., 2006). Their ability to establish both AM and to nodulate with rhizobial bacteria (but see Doyle and Luckow, 2003) confers to legumes a highly adaptive potential for colonization of harsh environments, and ability to enrich soils of primary habitats with valuable nitrogen, a strong capacity to produce storage proteins of great importance for human food and animal feeding (Graham and Vance, 2003), and the ability to synthesize allelopathic secondary nitrogen-based metabolites for defense against herbivores and for interspecific competition. Accordingly, endosymbioses in legumes became the most studied model systems in order to dissect the molecular and cellular bases underlying these two types of symbioses: from the external perception of the bacterial signal by the plant partner, to the establishment of the symbiotic interaction, with the objective to reconstruct the so-called "symbiotic program" (see for review: Geurts and Bisseling, 2002; Riely et al., 2006; Pawlowski and Sprent, 2008; Bucher et al., 2009; and references therein). In soils, rhizobia release symbiotic signaling compounds —Nodulation-factors (NF), consisting of β 1-4 N-acetylglucosamine (chitin) backbones—; which are encoded by bacterial*nod* genes. NFs induce root hair deformation and cell division in regions of the root inner cortex, resulting in the formation of nodule primordia. Hair roots respond to NF with branching and curling, which results in an entrapment of the bacteria in the infection pocket, where is initiated an intracellular infection thread. Among the main plant loci involved in this cascade of interactions, it was only recently demonstrated that different sub-families of receptor protein kinases (RPK) are required in the symbiotic transduction pathway for the establishment of the symbiotic complex, from the perception of the external bacterial signal to the activation of symbiosis-related genes.

Investigations in the model legumes revealed that AM and nodulation symbiosis share a partially over-lapping endosymbiotic genetic program, with at least seven common symbiosis genes (SYM) detected in *Lotus japonicus* (Kirstner et al., 2005; and references therein), suggesting recruitment of components from the ancestral plant-fungal pathway into the plant-bacterial nodulation program (Kistner et al, 2002; Markmann et al., 2008a, 2008b). Among the latter, key orthologous symbiosis receptor-like kinase genes, required for successful plant–microbial interactions with fungi and/or rhizobial bacteria, were identified and characterized in different legume systems: LjSYMRK in *Lotus japonicus* (Stracke, 2002); DMI2 and MsNork in *Medicago truncatula* and *Medicago sativa*, respectively (Endre, 2002); *Pisum sativum* (PsSym19; (Stracke, 2002; Endre, 2002); and SrSymRK in *Sesbania rostrata* (Capoen et al., 2005). Very recently, it was shown that the SymRK gene is also required in non-legume symbiotic interactions such asthe *actinorhizal* symbiosis of the cucurbit *Datisca glomerata* with an actinobacteria of the genus *Frankia* (Markmann, 2008b), and the formation of arbuscular mycorhiza in the Fagales tree *Casuarina glauca* (Gherbi et al., 2008).

In Rosids, the membrane-anchored protein encoded by the SymRK gene has a characteristic and well conserved receptor-like kinase structure, including: a signal peptide

(SP); an extracellular (EC) domain containing a specific binding site which perceives and initiates transduction of the microbial signals; a transmembrane (TM) domain that induces the chemical signal inside the cell; and an intracellular protein kinase (PK) domain that triggers off a cascade of molecular and cellular responses leading to the formation of endosymbiosis structures (see e.g., Riely et al., 2006; Pawlowski and Sprent, 2008; Markmann,2008a; Giles et al., 2008; Bucher et al., 2009). While the PK domain is highly conserved, the EC domain displays a significant sequence variability, both among orders, among legume genera, and among species in the genus *Medicago* (De Mita et al., 2007; Gherbi et al., 2008; Markmann, 2008a). The active site of the EC domain is characterized by the presence of three leucine-rich repeat (LRR) regions, which are involved in the perception of a likely specific extracellular ligand (Jones and Jones, 1997). To date, the activity and theligand of LRRs have not been clearly determined. Although, the activity of the SymRK is vital in the early stages of the root-bacterial association process, it has been suggested that the signal perception is not very specific, and therefore that the EC domain may not be directly involved in symbiotic specificity (Radutoiu, 2003; Gherbi et al., 2008; Holsters, 2008; Markmann, 2008b). At the infrageneric level, *Medicago*was the only genus examined to investigate relationships between SymRK evolution and rhizobial-host specificity. The data supported adaptive evolution of the NORK (synonymof SymRK) EC domain among *Medicago* species, but positive selection could not be associated with rhizobial specificity (De Mita et al., 2006, 2007). Regarding the crucial, but enigmatic, role of this gene in the establishment of the initial root-rhizobial association, it is of interest to extend investigations ofSymRK's evolution to representative genera from other legume lineages. Here, we explore the genus *Lupinus* L. (Papilionoidae; Genistoid alliance), which exhibits interesting changes in rhizobial specificity among species.

Lupinus is a natural and diverse group comprising about 300 annual and perennial herbaceous species, as well as soft-woody shrubs and few small trees. This genus occurs in a wide range of ecogeographical conditions in both the Old World, where 12-13 annual species are native to the Mediterranean region and North Equatorial Africa, and the New World, where more than 90% of the lupines are found, from Alaska to Terra de Fuego (Dunn, 1984; Planchuello-Ravelo, 1984; Gladstones, 1998; and reference therein). Regarding the notorious complexity of the genus, increasing efforts have been developed during the last decades to understand the diversity, the distribution, and relationships between lupines (see for review: Gladstones, 1974, 1984, 1998; Aïnouche and Bayer, 1999). Previous and complementary molecular phylogenetic investigations using various nuclear and plastid sequences, have significantly increased our knowledge of the evolutionary history of lupines, providing an improved systematic framework for the inference of character evolution, adaptation and speciation processes (Käss and Wink, 1997; Aïnouche and Bayer, 1999; Aïnouche et al., 2004; Citerne et al., 2003; Ree et al., 2004; Hughes and Eastwood, 2006; Drummond, 2008). Accordingly, several distinct lineages were revealed in *Lupinus*, in general accordance with the geographical distribution of the species in the Old and the New World. Nevertheless, despite much has been accomplished to circumscribe evolutionary lines and groups in *Lupinus*, there are still some fuzzy relationships and uncertainties remaining unresolved both at the base and within the genus, and a need for additional investigations to provide more informative characters.

Among the interesting biological and ecological features characterizing the genus, the most remarkable is its worldwide diversification, which implies adaptation to various and very contrasted eco-geographical conditions, including adaptive processes to establish beneficial symbiotic relationships with nitrogen-fixing bacteria. *Lupinus*is nodulated by a large variety of strains from the genus *Bradyrhizobium*, including both fast- and slow-growing

strains (Howieson et al., 1998; Stepkowski et al., 2007; and references therein). Most lupine isolates identified to date in both the Old and the New World were related to *B. japonicum*and *B. canariense*, and very few to *B. liaoningense*or *B. elkanii* (Stepkowski et al., 2007). Stepkowski et al. also showed a noteworthy parallel between the geographic pattern of divergence of the lupine bradyrhizobial strains and the diversification of their host plants. Additionally, there are evidence from the above studies and personal unpublished observations, that not only some geographic lupine lineages are associated to particular groups of rhizobialstrains, but also that diversification within a geographic lineage may be accompanied by remarkable changes in symbiont specificity. Therefore we examined whether correlations could be established between SymRK evolution, lupine diversification,symbiont diversity,and changes in the rhizobial specificity.

In this work, we have isolated and sequenced a large part of the nucleotide sequence encoding the extracellular domain of SymRK in 30 *lupine* taxa and 2 outgroups from the genus *Ulex* (Genisteae), in order to: (1) characterize this region, its diversity and sequence variability within the genus *Lupinus*; (2) reconstruct the SymRK phylogeny and evaluate its utility for inference of lupine relationships and history; and (3) examine rhizobial diversity and specificity observed in the genus in the light of the SymRK phylogeny and in relation with its patterns of sequence divergence.

Material and Methods

Plant material Thirty-seven accessions belonging to 28 annual and perennial species and subspecies were used in this study. Almost all Mediterranean and African Old World lupines were included in our sampling. Given the wide diversity of the New World lupines, only a limited number of taxa was selected to represent the main geographical North and South American groups, following insights from previous studies in the genus (Käss and Wink, 1997; Aïnouche and Bayer, 1999; Aïnouche et al., 2003; Aïnouche et al., 2004; Hughes and Eastwood, 2006; Drummond, 2008). Taxa known or suspected to preferably establish symbiotic associations with specific nitrogen-fixing rhizobia were represented: *L. princei*, endemic to East Africa (Kenya); *L. villosus*, representative of the singular simple-leaved lupines from southeastern US (Florida); and *L. paraguariensis*, representative of another singular group of simple and compound-leaved taxa native from Brazil, Paraguay and North Argentina (and related the Texan lupines). Additionally, two *Ulex* species (*U. parviflorus* Pourr. and *U. australis* Clemente) were chosen as outgroup for phylogenetic analysis. The genus *Ulex* is part of the Genisteae *sensu tricto*, the sister-group of *Lupinus* (Aïnouche et al., 2003, 2004; and references therein).

Accession numbers, geographic origins, and distribution of samples used in this study are summarized in Table 6–5. Almost all well identified taxa studied here were grown in the greenhouse at the University of Rennes-1 (France) to generate fresh material for molecular analyses. A few sequences were obtained from herbarium samples. Total DNA was isolated from 50-100 mg fresh leaf tissue with the Nucleospin Plant kit (Macherey-Nagel), following the manufacturer's instructions.

Primer design and amplification of the extracellular SymRK domain The SymRK region encoding the extracellular domain was specially targeted. This region corresponds to approximately 3,250 bp, and includes 6 exons in *Lotus japonicus*, (Figure 6.4 A and B). In order to design primers in highly conserved regions and to amplify this domain in *Lupinus*, the SymRK mRNA sequence available in GenBank for *L. albus* (AY935267) was compared to the complete SymRK sequence of *Lotus japonicus* genome (CM177) and to mRNA sequences from other Fabaceae such as: *Medicago truncatula* (AJ418369)*; Pisum sativum* (AF491997), *Astragalus sinicus* (AY946203); *Melilotus albus* (AJ428991); and *Sesbania rostata* (AY751547).

A series of tests allowed to select four couples of primers (F, forward; and R, reverse) covering the region encoding the extracellular part of the SymRK gene: F3-R5, F5c-R4a, F5-R3, and F5b-R3a (Table 6–6; Figure 6.4 C). All amplifications were carried out in a 50 μl reaction mixture containing 10 µl of 5× buffer, 0.1 µg of $MgCl_2$ (final concentration of 2 mM), 5 µl of dNTPs at 2 mM/each, 3 µl of each 5 µM primers, 1.25 unit of GoTaq polymerase (Promega GreenTaq), 10-50 ng of DNA. The PCR amplification starts with 3 min of DNA denaturation at 94°C, followed by 30-35 cycles of 30 s at 94°C, 60 s at 54°C and 60 s at 72°C for each cycle. A 7 min final extension at 72°C followed cycle 35. Double-stranded PCR products were purified with the EZ-10 Spin Column PCR Products Purification Kit (Bio Basic Inc.) and sent to Macrogen Inc. (Seoul, South Korea) for direct sequencing using the ABI system.

Cloning, sequencing and sequence alignment Most purified PCR products were subjected to cloning prior to sequencing (1 to 10 clones per sample) in order to evaluate sequence homogeneity and orthology of SymRK within *Lupinus*. Cloning reactions, isolation and purification of positive plasmids were performed using the pGEM-T easy Vector System and the Wizard Plus Minipreps DNA purification System II (Promega), following the manufacturer's instructions. Positive plasmids containing an insert were sequenced employing T7 and SP6 primers. Non-cloned products were directly sequenced both strands with their specific primers. Sequencing was performed by Macrogen Inc (Seoul, South Korea) using an ABI automated sequencer (Applied Biosystem). Raw sequences were cleaned from low quality regions and vector fragments with the software LUCY (Chou & Holmes, 2001 and Li & Chou, 2004). Cleaned fragments were assembled with CAP3 (Huang & Madan, 1999) and visualized with BIOEDIT 7 (Hall, 2005). Multiple alignment step was performed with the software M-COFFEE (Wallace et al., 2006; Moretti et al., 2007).

Orthology assessment Sequences available from GenBank and the literature show that *SymRK* is present as one copy in almost all legumes surveyed (except the two copies found in the tetraploid *Glycine max*), and that these copies are very likely orthologous (Endre et al., 2002; De Mita et al., 2006; Gherbi et al., 2008). As the SymRK was not previously studied in *Lupinus* (except one EST of *L. albus* available in GenBank), and regarding the paleopolyploid status of the genus (Dunn and Gillet, 1966; Fernandes and Queiros, 1978; Plitmann and Pazy, 1884; Gladstones, 1998; Maciel and Schifino-Wittmann, 2002; Canterato and Schifino-Wittmann, 2002; Camillo et al., 2006), the orthologous or paralogous nature of SymRK sequences had to be assessed. With this end in view, a set of randomly cloned sequences was search for eventual paralogous copies by phylogenetic analysis and evaluation of sequence similarity. Thus, lupine sequences were each verified to match the legume *SymRK* sequences *via* BLAST searches in GenBank (Altschul et al., 1990). A phylogenetic analysis was also performed on a reduced data set of *SymRK* coding sequences, including some lupine species and representatives of model and non-model legumes and Rosids, to verify their orthology. Moreover, predicted amino acid sequences of the EC domain of lupines were checked for the presence of leucine-rich repeats motifs (LRR), and compared to putative orthologous proteins from other legumes through a multiple alignment (M-Coffee).

Phylogenetic Analyses Phylogenetic analyses of the *SymRK* gene in *Lupinus* were based on a region of about 1500 bp generated for the 28 taxa used in this study. This region covers a large part of the extracellular domain, from exon 2 to exon 7. Sequence characteristics and pairwise sequence distances were calculated from the multiple alignment of the sequences using PAUP* 4.0b10 (Swofford, 2003).

Parsimony analyses were conducted using PAUP with heuristic searches and default options. Insertion-deletion events were

coded with the software SEQSTATE (Müller, 2005, 2006) using the Multiple Complex Index Coding (MCIC) method (Simmons and Ochoterena, 2000). The bootstrap method—1,000 replicates with heuristic search—was employed to estimate the robustness of clades (Felsenstein, 1985). For parametric analysis, the appropriate nucleotide substitution model of sequence evolution was determined by using JMODELTEST (Posada & Crandall, 1998; Posada, 2008, 2009) and the corrected Akaike information criterion (Posada & Buckley, 2004). The matrix and the selected model were then passed to PHYML 3 for analysis (optimisation of all parameters, 100 random starting points, and 1,000 bootstrap replicates). *SymRK* To infer divergence dates, a bayesian analysis was performed with the software BEAST (Drummond & Rambaut, 2007) using two DNA substitution models: Hasegawa-Kishino-Yano (HKY85) model and the General Time Reversible (GTR) model, both with a site heterogeneity model (Gamma distribution). The Markov chain Monte Carlo (MCMC) algorithm was set to run for 10 millions generations, under a relaxed clock model, with a burn-in of 10% and a sampling every 1,000 generations. For each model, five independent runs were realized from random starting points and logs were submitted to TRACER v1.4.1 (http://tree.bio.ed.ac.uk/software/tracer/) to confirm likelihood stationary, adequate mixing of the MCMC chains and convergence of independent runs. Time calibration of nodes in the resulting Bayesian *SymRK* tree was performed using 16.01 million years (Myr) as the age of the divergence between *Lupinus* and *Spartium* calculated by Hughes and Eastwood (2006) on the basis of data from previous fossil-calibrated molecular estimates in the legume family (Lavin et al., 2005). *Ulex* (the outgroup used here) and *Spartium* are both members of the Genisteae *sensu stricto*,which is sister group to *Lupinus* (Aïnouche et al., 2003; and reference therein). Accordingly, the same age of divergence can be used for *Ulex* and *Lupinus*.

Sequence variability and evolution Patterns of molecular evolution of the EC *SymRK* sequences have been examined among lupines, with particular attention to species showing differential rhizobial associations. For that purpose, synonymous (Ka) and non-synonymous (Ks) substitution rates were estimated and compared (Ka/Ks) to see whether or not adaptive evolution might have occurred in certain lineages (Messier and Stewart, 1997; Yang and Belawski, 2000). Accordingly, Ka/Ks ratios near 1 rather reflect neutral sequence evolution, Ka/Ks ratios less than 1 are considered as evidence of sequence evolution under negative or purifying selection, corresponding to high selective constraints, whereas Ka/Ks values greater than 1 is considered as evidence of sequence evolution under positive selection or relaxed selective constraints, generally suggesting adaptive evolution. After removing identical sequences to avoid division by zero, Ka/Ks ratios where calculated by the software KaKs_Calculator (Zhang et al., 2006), using maximum-likelihood methods and 19 different models of evolution. As stated by the authors, model averaging can reduce biases arising from model selection. We therefore chose the model-averaged method offered by KaKs_Calculator. In complement to that approach based on "pairs of taxa", codon-based models of molecular evolution are able to infer signatures of selection from multiple alignments of homologous sequences by estimating the rates of synonymous and non-synonymous substitutions (Poon *et al.*, 2009). Lupine *SymRK* sequences were submitted to three of those methods (SLAC, FEL and REL), embedded in the webservice Datamonkey (`http://www.datamonkey.org/`, Kosakovsky Pond & Frost, 2005).

Results

Isolation and characterization of the *SymRK* EC sequence in *Lupinus* The extracellular domain of *SymRK* was first amplified in two contiguous parts with primer couples F3- R5 and F5-R3 (Figure 6.4C). PCR

amplification of each part yielded a single band in agarose gel electrophoresis, showing no detectable evidence of size variation among lupine sequences (Figure not shown). Sequencing was performed with both PCR and intermediate primers (Table 6–6). Sequence identification was assessed by comparison with legume *SymRK* sequences available in GenBank via BLAST search algorithm. A preliminary phylogenetic analysis of random clones, obtained from five lupine species displaying various chromosome numbers, showed that all sequences grouped together according to their taxonomic origin (Figure 6.3)F. This topology is in agreement with species relationships known from previous lupine phylogenies (Aïnouche et al., 2004), which supports *SymRK* as a most likely single copy gene and that all sequences generated in this study may be regarded as putative orthologs. As no significant sequence heterogeneity and no paralogs were found among cloned sequences, each taxa is only represented by one clone in further analyses.

Intron/exon boundaries in the *Lupinus* EC *SymRK* domain were predicted by comparison of a reduced set of five complete aligned sequences, representing both Old and New world species, to the corresponding *SymRK* sequence of the model plant *Lotus japonicus* (Lj*SymRK* ; Figures 6.4B). As can be seen from both Figure 6.42C and Table 6–7, sequence comparison of *Lupinus* and *Lotus* revealed a highly conserved intron/exon microstructure of the *SymRK* EC domain. The region extending from exon 2 to the beginning of exon 9 corresponds to the EC domain, whereas the TM domain spans exons 9 and 10. Except intron 5, which showed a similar size (103-104 bp), all introns displayed variable sizes both among lupines and between *Lupinus* and *Lotus*, with 21.56 to 39.51 % of intron pairwise sequence divergence between the latter two. Apart exon 2, which was incompletely sequenced here, and exon 4, which showed 9 nucleotides less in *Lupinus*, the exons 3, 5, 6, 7 and 8 displayed identical sizes, with significant levels of mean sequence similarity ranging from 78.85 to 83.33 % (mean = 81.09 %). A phylogenetic tree of the *SymRK* coding sequence (exons 2 to 8) from five lupine species and four legume representatives is presented in Figure 6.5. This topology is congruent with previous legume phylogenies inferred from plastid DNA sequences (Kajita et al., 2001; Wojciechowski et al., 2004), and clearly shows a monophyletic origin for lupine sequences supporting orthology of sequences. Comparison of the proteins predicted from the same coding sequences allowed identification of three well conserved LRR motifs in the EC domain (overlapping exons 4, 5, 6, 7 and 8), providing additional support to *SymRK* sequence homology at the functional level (ref; Eisen, 1998).

Sequence variation and phylogeny of the *SymRK* ECD in *Lupinus* A data set of sequences covering the region from exon 2 to the end of intron 3 was generated for 38 lupine samples and 2 outgroups (Table 6–5), and was used for phylogenetic analysis.

Nucleotide pairwise-divergence among lupine sequences ranged from 0 to 7.99 %, with a mean-value of 4.66 %. Alignment of lupine sequences (1410 to 1479 bp) yielded a data matrix of 1,508 aligned characters (including 16 indels of 2 to 58 bp length, and several single gaps) of which 217 were parsimony informative. This number increased to 233 when indels were coded. The two *Ulex* species retained as outgroup exhibited several indels amongst which a remarkable 414 bp insertion in intron 2 increasing the overall size of the aligned-matrix to 1,967 bp and the number of parsimony-informative characters to 292 (coded indels included). Maximum parsimony (MP) analysis of the data set yielded 84 equally most-parsimonious trees of 722 steps (consistency index = 0.806, retention index = 0.889). A strict consensus tree with bootstrap support values was generated (not shown).

For parametric analysis, the model of evolution selected by jModelTest was TIM3+G (-lnL= 7510.9380 and weight = 0.2871). It is a "transitional model" allowing unequal base frequencies, four different substitution rates,

and rate variation among sites (Posada, 2008). In the model averaged phylogeny calculated by jModelTest, nodes appeared resilient to changes of selection model (Fig. not shown), indicating that they are stable to phylogenetic uncertainty due to model selection (Posada, 2009). This averaged phylogeny was identical to the topology produced by the ML analysis performed with PhyML, based on the TIM3+G model. The 100 random starting points analyses all converged to a narrow range of log-likelihood (-7510.67923 to -7510.68439), slightly better than the value obtained by jModelTest. These results suggest that the likelihood surface for the *SymRK* data set is dominated by one peak (Guindon et al., 2009). The data set led to only one general topology. The best topology is presented in Figure 6.6, with bootstrap values indicated only for nodes receiving more than 50% of support. This topology was in general accordance with the topology generated by the maximum parsimony method, with however more resolution and support in the ML tree. The main difference between the two topologies concerned the conflicting placement of *L. villosus*, and the lack of resolution for relationships between the three clades a, b and c in the New World assemblage IB (see details below).

Apart a few moderately to weakly supported relationships, most clades received strong bootstrap support in the ML analysis (Figure 6.6). All *SymRK* sequences here analyzed appear to have derived from a common ancestor (100% of bootstrap value) and to have evolved, since the origin of the genus, as two main lineages: one corresponding to the New World lupine sequences (hereafter called clade I), and the other to the Old World ones (hereafter called clade II). If the Old World clade received a support of 73.4%, the New World clade was only weakly supported by a bootstrap value of 56.7%. This Old World – New World disjunction was also observed in the MP topology, but with the exception of the sequence of *L. villosus* (Florida), which was either placed at the base of the New World sequences in the ML tree or placed at the base of the genus as sister to all other Old and New World lupine sequences in the MP tree, with however a weak support in either cases. According to the model-averaged phylogeny inferred by jModelTest (only one general topology found), the ML hypothesis is robust and not model-dependant, and is also supported by the bayesian analysis (presented later; Figure 6.7).

Therefore, in the New World clade I, the sequence of *L. villosus* (labelled IA), representative of the southeastern North-American unifoliolate lupines, was placed as sister to all other New World lupines (labeled IB), which are strongly supported as a monophyletic group (98/93% bootstrap, in the ML/MP trees, respectively). Within clade IB, three robust sub-clades (labeled IBa, IBb, and IBc) are identified with 97 to 100% support in either the ML or the MP analysis. Sub-clade IBa includes sequences from annual western North-American lupines referred to as the taxonomic group named *Platycarpos*.

The robustness of this group is highly reinforced by three shared indels of 6, 18 and 58 bp. Sub-clade IBb consists of sequences from perennial and annual species representative of western North American lupines (*L. arboreus*, *L. polyphyllus*), Central American (*L. mexicanus*, *L. elegans*), and western South American (*L. mutabilis*). Within the latter sub-clade, the *SymRK* data support sister relationships between Mexican and Andean lupines (83.1/77% bootstrap in ML/MP trees), which appear to derive from a western North American lineage. Sub-clade IBc clustered *L. texensis*, a representative of the digitate-leaved lupine endemics from southern North America (Texas), to the assemblage of all sequences originating from eastern South-American lupines (92.8/70% ML/MP support) with digitate (*L. magnistipulatus*, *L. bracteolaris* and *L. gibertianus*) or mixed digitate and unifoliolate leaves (*L. paraguariensis*). The placement of the eastern South-American lupines suggests that they derive from a North American lineage. According to the ML topology, sub-clades IAb and IAc are moderately supported as sister groups

(70.5% bootstrap), but this relation was not supported in the MP phylogeny where relationships among sub-clades a, b and c remained unresolved.

In clade II, the circum-Mediterranean smooth-seeded species were robustly resolved into four lineages corresponding to sections *Angustifoli* (IIA, including *L. angustifolius*), *Lutei* (IIB, incl. the sister species *L. luteus* and *L. hispanicus*), *Albi* (IIC, incl. *L. albus* and its synonymous *L. vavilovii*) and *Micranthi* (IID, incl. *L. micranthus*). The remaining Old World lupines, mainly occurring in North Africa and eastern Mediterranean, formed the section *Scabrispermae*, a strongly supported clade of species (IIE, 100% bootstrap, with one shared indel of 7bp), characterized by a typical rough seed coat microstructure. Within the latter section, variation of the *SymRK* sequences allowed circumscription of some well to moderately supported groups such as: a "*pilosus* group" (incl. *L. pilosus*, *L. palaestinus*, *L. anatolicus*) with 94.1/69% ML/MP bootstrap; a "*cosentinii-atlanticus* group" (55.1/55% ML/MP bootstrap); a "*digitatus-princei* group" (73/73% ML/MP bootstrap). Sister relationships between the latter two groups are supported by 72.9/64% ML/MP of bootstrap values. In both the ML tree and the strict consensus MP tree, the five OW sections were divided into two weakly supported groups: one including sections *Angustifoli* and *Lutei* (55.6/less than 50% ML/MP bootstrap); and the other including sections *Micranthi*, *Albi* and *Scabrispermae*. Section *Micranthi* being placed as sister to the well supported *Albi+Scabrispermae* clade (81.9/75% ML/MP bootstrap, with two shared indels).

Therefore, in the New World clade I, the sequence of *L. villosus* (labelled IA), representative of the southeastern North-American unifoliolate lupines, was placed as sister to all other New World lupines (labeled IB), which are strongly supported as a monophyletic group (98/93% bootstrap, in the ML/MP trees, respectively). Within clade IB, three robust sub-clades (labeled IBa, IBb, and IBc) are identified with 97 to 100% support in either the ML or the MP analysis. Sub-clade IBa includes sequences from annual western North-American lupines referred to as the taxonomic group named *Platycarpos*. The robustness of this group is highly reinforced by three shared indels of 6, 18 and 58 bp. Sub-clade IBb consists of sequences from perennial and annual species representative of western North American lupines (*L. arboreus*, *L. polyphyllus*), Central American (*L. mexicanus*, *L. elegans*), and western South American (*L. mutabilis*). Within the latter sub-clade, the ***SymRK*** data support sister relationships between Mexican and Andean lupines (83.1/77% bootstrap in ML/MP trees), which appear to derive from a western North American lineage. Sub-clade IBc clustered *L. texensis*, a representative of the digitate-leaved lupine endemics from southern North America (Texas), to the assemblage of all sequences originating from eastern South-American lupines (92.8/70% ML/MP support) with digitate (*L. magnistipulatus*, *L. bracteolaris* and *L. gibertianus*) or mixed digitate and unifoliolate leaves (*L. paraguariensis*). The placement of the eastern South-American lupines suggests that they derive from a North American lineage. According to the ML topology, sub-clades IAb and IAc are moderately supported as sister groups (70.5% bootstrap), but this relation was not supported in the MP phylogeny where relationships among sub-clades a, b and c remained unresolved.

In clade II, the circum-Mediterranean smooth-seeded species were robustly resolved into four lineages corresponding to sections *Angustifoli* (IIA, including *L. angustifolius*), *Lutei* (IIB, incl. the sister species *L. luteus* and *L. hispanicus*), *Albi* (IIC, incl. *L. albus* and its synonymous *L. vavilovii*) and *Micranthi* (IID, incl. *L. micranthus*). The remaining Old World lupines, mainly occurring in North Africa and eastern Mediterranean, formed the section *Scabrispermae*, a strongly supported clade of species (IIE, 100% bootstrap, with one shared indel of 7bp), characterized by a typical rough seed coat microstructure. Within the latter section, variation of the *SymRK* sequences allowed cir-

cumscription of some well to moderately supported groups such as: a "*pilosus* group" (incl. *L. pilosus*, *L. palaestinus*, *L. anatolicus*) with 94.1/69% ML/MP bootstrap; a "*cosentinii-atlanticus* group" (55.1/55% ML/MP bootstrap); a "*digitatus-princei* group" (73/73% ML/MP bootstrap). Sister relationships between the latter two groups are supported by 72.9/64% ML/MP of bootstrap values. In both the ML tree and the strict consensus MP tree, the five OW sections were divided into two weakly supported groups: one including sections *Angustifoli* and *Lutei* (55.6/less than 50% ML/MP bootstrap); and the other including sections *Micranthi*, *Albi* and *Scabrispermae*. Section *Micranthi* being placed as sister to the well supported *Albi+Scabrispermae* clade (81.9/75% ML/MP bootstrap, with two shared indels).

Time calibration of nodes in the *SymRK* phylogeny Both independent runs performed with HKY85 and GTR models converged to the same range of log-likelihoods—between 7535 and 7555—comparable to the log-likelihood obtained with PhyML. The GTR model was the best fitted to our data set. Accordingly, the GTR-based phylogeny of the *SymRK* EC domain is presented in Figure 6.7. This topology is highly congruent with that derived from the maximum likelihood analysis (Figure 6.6). Almost all nodes had strong support with posterior probabilities comprised between 0.87 and 1.0. Using 16.01 million years (Myr) as the age of the stem node of *Lupinus* (Hughes and Eastwood, 2006) to calibrate nodes of the *SymRK* tree (Figure 6.7), the main divergence at the base of the genus between the New (I) and the Old World (II) lineages is estimated at ca. 6.66 Myr ago during the Miocene. Most stem nodes of the major clades observed in both regions appeared during late Miocene and Pliocene (6-2 Myr), with however some significant differences amongst them. In the New World, *L. villosus* (IA) appears to derive from an early divergent line (6.08 Myr). The most recent common ancestor of all the remaining American lupines (IB) began to diversify ca. 4.66 Myr and generated novel lines (a, b, c), which in turn experienced diversification events at the end of the Pliocene and beginning of the Pleistocene (ca. 2.37 Myr in IBc; ca. 2.03 Myr in IBc; ca. 1.52 Myr in IBb). Additionally, estimated divergence times within clade IB indicated that only recently lupines diversified in Mexico and the Andes (ca. 1.21 Myr) and in eastern South America (ca. 1.05 Myr) following independent colonization events from separate North American lines. In the Old World, estimated ages of stem nodes leading to extant lupine sections were between ca. 5.53-4.32 Myr. Within section *Lutei*, *L. luteus* and *L. hispanicus* diverged during mid-Pliocene (ca. 2.93 Myr), whereas *Scabrispermae* appeared to have more recently diversified into two groups ca. 1.47-1.19 Myr (Pleistocene) in the Mediterranean region and North Africa, after theydiverged from their common ancestor (ca. 1.72 Myr).

Detecting signatures of selection in the *SymRK* EC domain of *Lupinus* The main lines of known symbiotic *Bradyrhizobium* strains previously reported to effectively nodulate lupine roots are indicated in the *SymRK* phylogeny (Figure 6.6) according to data from Stepkowski et al. (2007). Most rhizobial strains isolated from the Old World lupines, such as *L. albus* and *L. angustifolius*, belong to *B. canariensis* (BC). In the New World, lupines such as *L. mutabilis*, *L. arboreus*, *L. polyphyllus* and *L. bracteolaris* are mainly nodulated by *B. japonicum* (BJ) strains, which also were punctually detected in *L. albus*. Whereas only a strain from *B. elkanii* (BE) was reported so far for the eastern South American species *L. paraguariensis* (South Brazil). Additionally to this rhizobial shift observed for *L. paraguariensis* in clade IBc, also it is obvious that other particular lupine species showed specific rhizobial requirements, such as for example: *L. princei* from Kenya in North Equatorial Africa, as compared to its close relatives in the *Scabrispermae* (Howieson, 1998); and the unifoliolate-leaved lupine native from the sub-tropical area of south-eastern USA (in Florida). Seedlings of the latter species never

reached a fully developed stage and never flowered when they were grown out of their native soil in our greenhouse, where most other lupines did (excepted *L. princei* and *L. paraguariensis*; unpublished personal observations). Thus, to evaluate the potential effect of selection on the EC domain of *SymRK* within *Lupinus*, and most particularly to examine whether positive selection occurred in the lupine lines which seem to have experienced a shift in rhizobial-specificity (branches are labeled in the phylogeny Figure 6.6), we performed a broad interspecific analysis of pairwise sequence comparisons using Ka/Ks ratios, complemented by a codon-based analysis.

Ka/Ks comparison of our five longest available lupine sequences (exons 2 to 8) with a selection of four *SymRK* mRNAs issued from model legumes showed a dominant signal of purifying selection (<1), with values ranging from 0.36 to 0.52 (average 0.45, see Table 6.8A). Comparison among a broadest selection of lupines based on a region including exons 2 to 6 yielded a similar general pattern characterized by a dominant signal of negative selection (Ka/Ks values ranging from 0.15 to 1.33; 0.45 on average), with 99% of Ka/Ks ratios <1 (Table 6.8B). Two extreme values, not taken into account in the preceding range, are due to a very low number of substitutions. First, the Ka/Ks ratio between *L. digitatus* (T4) and *L. princei* (T17), two closely related and poorly divergent African species (Figure 6.5), is close to zero with only two synonymous substitutions between the two species. Secondly, the Ka/Ks ratio between *L. microcarpus* (T27) and *L. luteolus* (LT), both members of *Platycarpos*, is 41.82 with only one non-synonymous substitution between the two species (models deal with real numbers not integers, which explains the absence of null or infinite values). The positive selection (ratio 1.33) detected between the closely related *Scabrismermae L. palaetinus* (T14) and *L. pilosus* (T6) is also due to a limited number of substitutions (11) and is clearly not statistically supported (p-value 0.97). Although almost all pairwise comparisons of *SymRK* EC domain gave Ka/Ks values <1, of which 68% were < 0.5, it is interesting to notice that 30% were between 0.5 and 1. This is particularly observed for two species: *L. micranthus* (T19) from the Mediterranean region; and *L. villosus* (K36), the southeastern US lupine showing specific rhizobial requirements. The latter species exhibited several Ka/Ks values close or equal to 1, suggesting a trend toward a neutral sequence evolution. *Lupinus micranthus* displayed the highest values when compared to the *Scabrispermae* (0.73 to 1.00), to section *Albi* (0.75), to the *Platycarpos* (0.82 to 0.87), to section *Angustifoli* (0.76–0.87) and to *L. villosus* (0.94). The latter also showed a relative increase of its Ka/Ks values relative to the *Scabrispermae* (0.63 to 0.83). These long-branched species, *L. micranthus* and *L. villosus* (Figure 6.6) appeared to be of old origin in the genus (ca. 6.08-5.86 Myr) and exhibited very divergent sequences from those of their congeners. However, most of their highest Ka/Ks ratios were not supported by statistical tests based on the present data set (*e.g.*, with p values generally > 0.5). Also two other Ka/Ks values were remarkable, but both were not statistically supported, due to very low number of substitutions: 0.89 ($p < 0.62$) between the two *L. hispanicus* subspecies (*hispanicus* and *bicolor*) in the Mediterranean sect. *Lutei* (clade IIB); and 0.79 ($p < 0.66$) between *L. paraguariensis*—,which showed specific rhizobial *B. elkanii* strains—, and *L. bracteolaris*, which is nodulated by *B. japonicus* strains (more ubiquitous in the New World clade IB).

As no sign of branch-specific positive selection was detected, we tried to detect position-specific signatures of selection in the lupine ***SymRK*** EC domain using a codon-based analysis. Sequences coding for proteins implicated in signal recognition, like ***SymRK***, can present positive selection. The tools packed in the Datamonkey software suite have been used on the matrix of lupine sequences including exons 2 to 6 (402 codons). Negatively or positively selected positions detected by SLAC, FEL and REL methods have been integrated in Figure 6.9, with positions

indicated by at least two of the three methods topped by a balloon. A total of 42 negatively selected positions have been detected (in blue on Figure 6.9), from which 11 are supported by at least two methods. Negatively selected positions are evenly dispersed along the sequence, with no particular pattern. Similarly, amino acids encoded by negatively selected positions are evenly distributed in the Venn's diagram (not shown) of amino-acid properties (Livingstone & Barton, 1993), indicating that hydrophilic or hydrophobic properties do not seem preferentially selected. Only two positions are reported as positively selected from both two methods (in red in Figure 6.9), and none of them was localized in the active LRR region. A similar analysis performed on a shorter matrix covering exon 2 and 3 (307 codons) shows a greater number of negatively selected positions (81 positions, 29 multi-supported).

Discussion

In this work, we have isolated and sequenced a large part of the DNA sequence encoding the extracellular domain of *SymRK* gene in 30 *lupine* taxa (38 accessions) and 2 outgroups from the genus *Ulex* (Genisteae), in order to: characterize this region and its sequence variability within the genus *Lupinus*; reconstruct the *SymRK* phylogeny and evaluate its utility for inference of the evolutionary history of the lupines; and examine whether correlations can be established between the molecular pattern of divergence of the *SymRK* EC domain and the rhizobial diversity and specificity observed in the genus.

Structure and orthology assessment of the *SymRK* gene in *Lupinus* Since the discovery of the symbiotic receptor kinase gene in *Lotus japonicus* (Lj*SymRK*; Stracke et al., 2002) and *Medicago truncatula* and *M. sativa* (DMI2 and NORK; Endre et al., 2002), only a single copy was generally isolated from each of the legume taxa surveyed for this gene. Regarding their high level of similarity in terms of structure, sequence, function and genomic location, all these copies were regarded as homologous sequences in the legume family (Endre et al., 2002; Zhu et al., 2005). Moreover, phylogenetic analysis of these copies support their descent from a common ancestral copy (e.g., Gherbi et al., 2008) and generated a topology which is consistent with previous legume phylogenies inferred from plastid DNA sequences (Kajita et al., 2001; Wojciechowski et al., 2004). This is illustrated in this study (Figure 6.5) using a reduced data set. Among the legumes surveyed to date (available in GenBank), only *Glycine max* exhibited two weakly divergent copies (92.4% of identity), likely resulting from a recent specific round of polyploidy or segmental duplication (Pfeil et al., 2005; Shoemaker et al., 2006; and references therein). Within *Lupinus*, all sequences of the *SymRK* EC domain generated in this study, *via* either cloning and sequencing or direct sequencing of PCR products, showed a high level of similarity (maximum of pairwise divergence: 8%), and their phylogenetic analysis did not reveal any evidence or clue of paralogs among species. Comparison of the *SymRK* EC domain of *Lupinus* (a member of the Genistoid alliance) to those available for model and crop legumes (representative of other Papilionoid lineages), shows the high conservation of this gene among different legume lineages in terms of intron/exon microstructure, sequence similarity, and conservation of the putative functional LRR motifs. Additionally, phylogenetic analysis of coding sequences of the lupine *SymRK* EC domain in the Papilionoid context clearly shows a monophyletic origin for lupine sequences. All together the data strongly suggested that all lupine *SymRK* sequences generated here (and those of the outgroup *Ulex*) are orthologs of a single copy gene, and most likely represent putative homologues of *SymRK* sequences previously identified in legumes. Therefore, the data suggest that only one orthologous copy of *SymRK* would have been maintained over a long evolutionary time in papilionoid legumes. This is noteworthy, regarding that, as most angiosperms, the papilionoid legume

genomes (including the genistoids) also have experienced various evolutionary processes along their diversification, such as chromosomal rearrangements, ancient and recent polyploidization events, segmental and gene duplications which have been accompanied by selective gene loss/retention or pseudogenization of duplicated genes (Doyle and Lukow, 2003; Zhu et al, 2005; Pfeil et al., 2005; Cronk et al., 2006). *Lupinus* itself is a complex aneuploid series regarded as a paleopolyploid genus which most likely also has experienced neopolyploid and duplication events and various molecular evolutionary processes, regarding its diverse chromosome numbers ranging from $2n$ = 32 to 52 (punctually 96, 100) chromosomes with $x = 6$ as a most probable basic chromosome number among diverse propositions (Dunn and Gillet, 1966; Fernandes and Queiros, 1978; Plitmann and Pazy, 1984; Gladstones, 1998; Maciel and Schifino-Wittmann, 2002; Conterato and Schifino-Wittmann, 2002; Camillo et al., 2006). Thus, there is evidence suggesting that strong evolutionary forces tend to retain one *SymRK* copy per genome at various taxonomic levels to preserve the key ability of legumes to establish root-microbes interactions, while duplicated copies would be rapidly eliminated or pseudogenized following relaxation of purifying selection (Ohno, 1970). This is in agreement with the dominant signal of purifying selection observed here in Ka/Ks-based pairwise comparisons (< 0.52) among the *SymRK* mRNAs of model legumes and lupines. A similar pattern of evolution has been described for the transcription factor *LEAFY* in Brassicaceae, which suggested that duplicate LFY genes have not persisted through multiple speciation events and were generally lost by drift in relation with inflorescence evolution (Baum et al., 2005).

Phylogeny of the *SymRK* and its utility in *Lupinus* Sequence analysis of a region of about 1,500 pb of *SymRK* EC domain (including intron 2, exon 2 and intron 3) for the first time in thirty *Lupinus* species generated a significant number of parsimony informative nucleotide substitutions (217). This number is much higher than those found in most nuclear and plastid regions previously analyzed in *Lupinus* (Käss and Wink, 1997; Aïnouche and Bayer, 1999; Aïnouche et al., 2003, 2004; Drummond, 2008), with the exception of the *LEGCYC1A* locus displaying few more informative characters (Ree et al., 2004; Hughes and Eastwood, 2006). Phylogenetic analysis of the *SymRK* EC domain using the ML method generated a robust general topology, which was in general accordance with the MP tree and highly consistent with the Bayesian tree. Despite a low resolution at the base of the genus, *SymRK* sequences of the extant lupine species appear to have evolved during the last 7-8 Myr into two main lineages in the Old and the New World. Within each of the Old and New World assemblages, the *SymRK* phylogeny allowed circumscription of several well-supported groups, in general accordance with the eco-geographic distribution of the species.

Accordingly, the data suggest two distinct and early diverged (ca. 6 Myr) lineages in the New World: one represented by a small group of singular unifoliolate lupines native from South-eastern North-America (here represented by *L. villosus*), which position will be discussed below; and the strongly supported lineage containing the majority of the other New World lupines (over 90% of the genus) which have remarkably diversified during Pliocene and Pleistocene in North, Central and South America. Within the latter, the ***SymRK*** data identified three robust subclades. Two of them include all representatives of the lupine occurring in the western areas of the New World from Alaska, Mexico to the Andes: one containing the annual lupines with sessile and connate cotyledons referred to as the *Platycarpos* species, which mostly diversified in western North America (since ca. 2 Myr); the other comprised representatives of the majority of annual and perennial lupines occurring in western North, Central and South America and which have likely derived from a common recent ancestor ca. 1.5 Myr. The third sub-clade included

representatives of two distant geographical groups which diverged ca. 2.37 Myr: the digitate-leaved lupines native from southern North America in Texas (e.g., *L. texensis*); and the recently diversified (< 1-1.5 Myr) central-eastern South American lupines, which have various leaf types (digitated, unifoliolate or mixed leaves). Although relationships among these three main New World lineages (exclusive of the unifoliolate North American lupines) are only moderately supported, there is strong evidence from ***SymRK*** which support that Central and South American lupines most likely derived from North American lupines, following two independent routes of diversification and colonization: one (ca. 1.5 Myr) towards Mexico and Central America, with a subsequent rapid and remarkable radiation (ca. 1.21 Myr) onward to western South America following the Andes uplift, as suggested by Hughes and Eastwood (2006) based on ITS+LEGCYC1A and supported by additional cpDNA data of Drummond (2008); and the other route, towards central-eastern South America (ca. 1.05 Myr) very likely *via* a long distance dispersal.

In the Old World assemblage, the *SymRK* data depicted the early diversification of the Mediterranean and African lupines (end of Miocene, beginning of the Pliocene) and clearly distinguished the main lineages traditionally recognized at the section and species level, such as: the circum-Mediterranean smooth-seeded sections *Angustifoli*, *Lutei*, *Albi* and *Micranthi*; and the mainly African rough-seeded section *Scabrispermae*. Despite a generally weak or moderate support, all methods used here (MP, ML and Bayesian) generated the same pattern of relationships among the main Old World lineages, grouping together: on one side, sections *Angustifoli* and *Lutei*; and on the other side, sections *Micranthi*, *Albi* and *Scabrispermae*, with a significant support to sister relationships between the latter two sections. This pattern suggests that most likely the *Scabrispermae* derived ca. 5 Myr from a Mediterranean lupine lineage. Within the latter section, *SymRK* sequences provided some new insights suggesting that the extant *Scabrispermae* would have recently diversified in two lineages (ca. 1.47-1.19 Myr), after they diverged from their more recent common ancestor (ca. 1.72 Myr). One lineage is represented by the *Pilosus* group occurring in the eastern Mediterranean region (Near-East, Anatolia); while the other correspond to the *Cosentini-Atlanticus-Digitatus* group occurring in arid and desert high lands of North Africa (from Morroco, Mauritania, Central Sahara, to Nil Valley in Egypt), from which derived very recently *L. princei* (< 0.2 Myr), the only Old World species which is endemic to Kenya and adapted to equatorial conditions.

Therefore, ***SymRK*** sequences exhibited a good potential for phylogenetic inference within *Lupinus*. ***SymRK*** data alone generated a topology which is more resolved than previous single gene based phylogenies (Käss and Wink, 1997; Aïnouche and Bayer, 1999; Aïnouche et al., 2004), and is broadly in agreement with more recent lupine phylogenetic estimates, using either combined ITS and *LEGCYC1A* nuclear sequences (Hughes and Eastwood, 2006) or a combination of three plastid DNA sequences (Drummond, 2008). However, although they provided significant additional support to several groups and lineages in the genus, and shed some additional light into relationships within the Old World lupines (e.g., within *Scabrispermae*, and among the latter and sect. *Albi*), ***SymRK*** sequences did not yield sufficient signal to elucidate some uncertainties and conflicts left unresolved by all previous phylogenies, especially at the base of the genus among the main Old and New World lineages. Apart from some weak differences among these phylogenies, the main incongruence observed concerns the placement of *L. villosus*, the representative of the unifoliolate North American lupines, previously identified as a monophyletic group by Hughes and Eastwood (2006). This lineage was always clearly distinguished from the other New and Old World clades. However, in this study, it was either unresolved at the base of the genus (in the MP analysis) or placed as sister to the broad New World clade (IB) in both the ML and Bayesian analy-

ses; whereas, it was rather placed among the Old World lineages, but at various positions, depending on the genes and methods used for phylogenetic inference (Hughes and Eastwood, 2006; Drummond, 2008). In either case, only the Bayesian method strongly supported its placement. Such genetic heterogeneity, which includes sequences from different origins (Old and New World), is suggestive of a possible early reticulate origin of the unifoliolate NA lupines, or of other recombinational biological processes, which are often involved in incongruent gene trees (Wendel and Doyle, 1998; Seelanan et al., 1997; Small et al., 2004; Alvarez and Wendel, 2003). Thus, additional data and investigations are required to both elucidate the origin and placement of this key lineage at the Old World/New World junction, and to more accurately resolve relationships at the base of the genus, especially among the Old World lineages. This is of great importance to fill the few remaining gaps, and hence to better understand the origin and the evolutionary processes that have accompanied the early diversification of the lupines.

Lupine/rhyzobial symbiotic association and *SymRK* evolution In this study we have examined the diversity of the *Bradyrhizobium* lines known to nodulate with lupines (Howieson, 1998; Stepkowski et al., 2007) in the light of the *SymRK* phylogeny. As previously shown by Stepkowski et al. (2007), the distribution of the Bradyrizobial lines is broadly correlated with their geographic and lineage host range. The Mediterranean smooth seeded OW lupine species are mainly infected by bacterial strains from *B. canariensis*, which are also capable to nodulate North-African rough-seeded species (*Scabrispermae*) such as for instance *L. cosentinii*. Among the *Scabrispermae*, *L. princei* from Kenya, which was shown to derive from a North African rough-seeded lineage (likely from the *L. digitatus* line), makes an exception (Howieson, 1998). This species seems to have developed a symbiotic specificity with other bacterial strains (not yet identified) along with its adaptation to novel ecological conditions, following its recent colonization of equatorial areas. Most bacterial strains isolated from members of the broad New World clade IB belonged to *B. japonicum*. The latter also occurs in the Old World where it nodulates with other *Genistea* (including some lupines). It has been suggested that most *B. japonicum* bacterial strains that infect American lupines have very likely a euro-Mediterranean origin (Stepkowski et al., 2007). Within this clade, strains from *B. elkanii* were only isolated from *L. paraguariensis*, a representative of the unifoliolate lupines deriving from within the central-eastern South American lineage (clade IB). These species occurs in highlands of subtropical areas in southern Brazil and adjacent regions. Thus, it appears that lupine species exhibiting strong symbiotic specificity, such as *L. princei* (Kenya), *L. villosus* (Florida), and *L. paraguariensis* (South Brazil), all represent independent lines that colonized tropical and sub-tropical areas in Africa and America, after they diverged from different Old World or New World temperate lineages. As suggested above for *L. princei*, their rhizobial shift is likely due to their adaptation to new bacterial strains that have diverged from temperate region strains.

Examination of whether such adaptive processes were accompanied by positive selection on the EC domain of *SymRK* within *Lupinus* was performed using both interspecific pairwise sequence comparisons of Ka/Ks ratios and by a codon-based analysis. Apart from few artifactual Ka/Ks values, our results revealed a dominant signal of negative selection, indicating that the EC domain of the *SymRK* is under a strong purifying selection in *Lupinus*. None of the branches concerned by rhizobial shift showed a significant increase of non-synonymous substitutions. It has been shown that 30 % of the Ka/Ks values were ranged between 0.5 and 1, particularly in pairwise comparisons among long branched lineages, such as *L. micranthus*, *L. villosus*, *Scabrispermae* and *Platycarpos*. This would suggest an increase of non-synonymous substitutions in some branches

and hence a somewhat relaxation of purifying selection towards the neutral evolution of the *SymRK* sequences. However, none of these values received a significant support from statistical tests. Codon-based analysis detected numerous positions under negative selection. These positions were scattered along the sequence encoding the *SymRK* EC domain with no obvious particular pattern of distribution, and appeared to have no preferential effects on hydrophilic or hydrophobic amino-acid properties of the encoded protein. Only two positions were under positive selection but none of them was localized in the LRR domain. LRR repeats are the only known motif identified on the extracellular part of *SymRK* (see Fig 6.4D). A LRR repeat is composed of a highly conserved core (LxxLxLxxNxL) followed by a more variable part (xxxLPxxL) (N can be replaced by S, L can be replaced by I or F). Three-dimensional structures encoded by LRR are implicated in protein-protein recognition (Kobe & Kajava, 2001) and are widely distributed in the tree of life. Receptor like-kinases (RLKs) are a vast family of proteins to which belong *SymRK*. Some of these RLKs are implicated in pathogen recognition. Therefore, some positions of their LRR repeats are under diversifying selection. No such situation appears in *SymRK* where no positively selected position is reported. As the strength of the signature of selection increases with the number of sequences sampled, future increases of our taxa sampling may produce a clearer image and a most accurate functional annotation of the *SymRK* sequence. Therefore, no evidence of branch-specific positive selection, neither significant position-specific signatures of selection were detected in the lupine *SymRK* domain.

In conclusion Only one copy of *SymRK* was detected in all lupine species surveyed in this study. The data indicate that this gene is likely subject to strong selective constraints to maintain only one copy per genome. Compared to the nuclear and plastid DNA sequences previously employed, the region encoding the extra-cellular domain of *SymRK* provided a good signal for phylogenetic inference in *Lupinus*. However, further investigations using additional informative characters are needed to elucidate the few uncertainties and conflicts left unresolved. Our data are consistent with a strong purifying selection operating on the *SymRK* in *Lupinus*, and no signatures of positive/diversifying selection was found in lineages which experienced changes in rhizobial specificity. Thus, although *SymRK* was demonstrated as a vital gene in the early stages of the root-bacterial symbiotic associations, there is no evidence from present analyses indicating that this gene is involved in rhizobial specificity in *Lupinus*.

References

(see p. 191)

TAB. 6–5 — List of *Lupinus* and outgroup taxa included in this study. Samples are presented with their origin, geographic distribution, reference number, and GenBank accession numbers of their *SymRK* sequences. Abbreviations: OW, Old World; NW, New World; NA, North America; SA, South America; CA, Central America; Afr, Africa; Med, Mediterranean. USDA, US Department of Agriculture, Washington; INAE-DZ, Institut National d'Agronomie, El-Harrach, Algérie; AKA, Abdelkader Aïnouche; INRAL, INRA, Lusignan, France; BRA, EMBRAPA, Brasil. RP, Rémy Pasquier; UFRGS, Universidade Federal do Rio Grande do Sul; IV, Institut Vavilov, St Petersburg; WADA, Western Australia Department of Agriculture.

Taxon	**2*n***	**Origin/Distribution**	**Sample source & Reference number**
L. affinis	48	Oregon/NW, West NA	USDA/504315/N20
L. albus	50	Algeria/OW, Med	INAE-DZ/M20
L. anatolicus	42	Turkey/OW, Afr	AKA/K32
L. angustifolius ssp. *reticulatus*	40	France/OW, Med	AKA/T25
L. angustifolius ssp. *angustifolius*	40	Algeria/OW, Med	AKA-M1/T24
L. arboreus	48	California/NW, NA	USDA/393932/N88
L. argenteus	48	Washington/NW, NA	USDA/504374/N23b
L. atlanticus	38	Morocco/OW, Afr	USDA/384612-FM83-T1
————	38	Morocco/OW, Afr	INRA-SAPF/T11
L. bracteolaris	32-34	Brazil/NW, South-East SA	USDA/404349/S80
L. cosentinii	32	?/OW, Med	INRAL-FR/T15
L. digitatus	36	Egypt/OW, Afr-Med	WADA-PI26877/K34
————	36	Egypt/OW, Afr-Med	WADA-PI26877/T4
L. elegans	48	Mexico/NW, West SA	USDA/185099/S33
L. gibertianus	36	Brazil/NW, East SA	UFRGS/1835MTSW
L. hispanicus ssp. *bicolor*	52	Spain/OW, Med	USDA/PI 384554/T23
L. hispanicus ssp. *hispanicus*	52	Portugal /OW, Med	USDA/384555/T22
L. luteolus	?	California/NW, West NA	USDA/284721/LT
L. luteus	52	Algeria/OW, Med	AKA/M5
————	52	Algeria/OW, Med	AKA/T20
————	52	Algeria/OW, Med	AKA/T21
L. magnistipulatus	36	Brazil/NW, East SA	UFRGS/1840 MTSW
L. mariae-josephi	52 ?	Spain/OW, Med	H. Pascual/MJ1
L. mexicanus	48	Mexico, F.D/NW, CA	USDA/14748/N51
L. micranthus	52	Algeria/OW, Med	AKA/T19
L. microcarpus	?	?/NW, NA	?/T27
L. mutabilis	48	Perou/NW, West SA	INAE-DZ/S35/MU35
L. palaestinus	42	Near-East/OW, Afr-Med	INRA-FR/T14
L. paraguariensis	36	Brazil/NW, East SA	BRA-02828/BZ1
L. pilosus	42	North-Africa/OW, Afr-Med	IV/T10
————	42	Algeria/OW, Afr-Med	INAE-DZ/T6
————	42	Algeria/OW, Afr-Med	INAE-DZ/T9
————	42	North-Africa/OW, Afr-Med	USDA/W6 PI 11995/T13
L. polyphyllus	48	USA/NW, NA	USDA/504404/T26
L. princei	38	Kenya/OW, Afr	WADA P 23021/T0
————	38	Kenya/OW, Afr	RP Chyulu 1800/T16
————	38	Kenya/OW, Afr	RP Chyulu 1915/T17
L. texensis	36	USA/NW, South NA	USDA/577291/N45
L. vavilovii	50	Yougoslavia/OW, Med	Russia K3118/VAV1
L. villosus	52	Florida/NW, South-East NA	D. Jones/K36
Ulex australis	?	Portugal/OW, Med	AKA/D27
Ulex parviflorus ssp. *parviflorus*	32	Spain/OW, Med	AKA/G24

TAB. 6–6 — Designed primers for amplification and sequencing of the external domain of the *SymRK* gene.

Primer	Sequence (5′–3′)	Melting Temp. (°C)
F3	CAATAGATCATAGCTGGTTCTCTG	61.1
F5	CTTAATCTAACCTTGGTCAAGGC	60.9
F5b	CCTCGAGTAATCTCAAAGGAAC	56.6
F5c	GTCCTACAAACAGCTCTTACT	58.6
R3	GCCAAAGTAATCAAGATTGATCCAC	61.3
R3a	TTGATTCTGGAAGTGACCCC	60.4
R3p	CAAGCAGTGCTTGTTGATCTGTC	59.2
R4a	GKATCACTTCCACTGAWGAAATG	60.1
R5	GCCTTGACCAAGGTTAGATTAAG	60.9
R5c	AGTAAGAGCTGTTTGTAGGAC	58.6

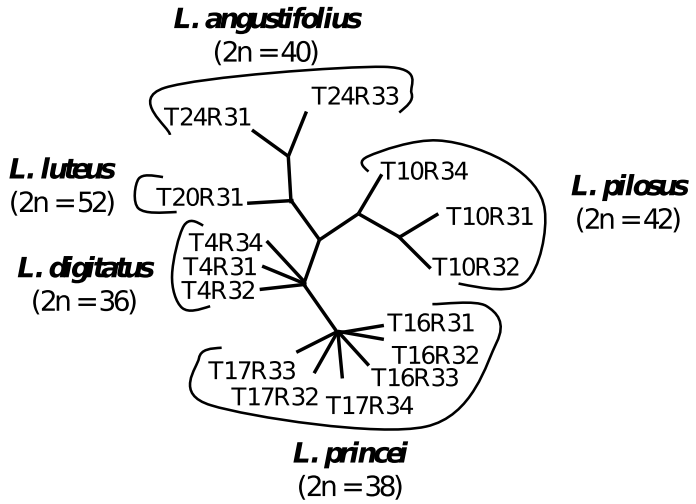

FIG. 6.3 — Phylogenetic analysis (maximum parsimony; tree length = 118 steps; CI = 0.983) of 15 random clones of a 1,274 bp fragment from the extracellular *SymRK* domain. There is no evidence of *SymRK* duplication in *Lupinus*. Clones from each species cluster together, and the topology is congruent with known lupine species trees.

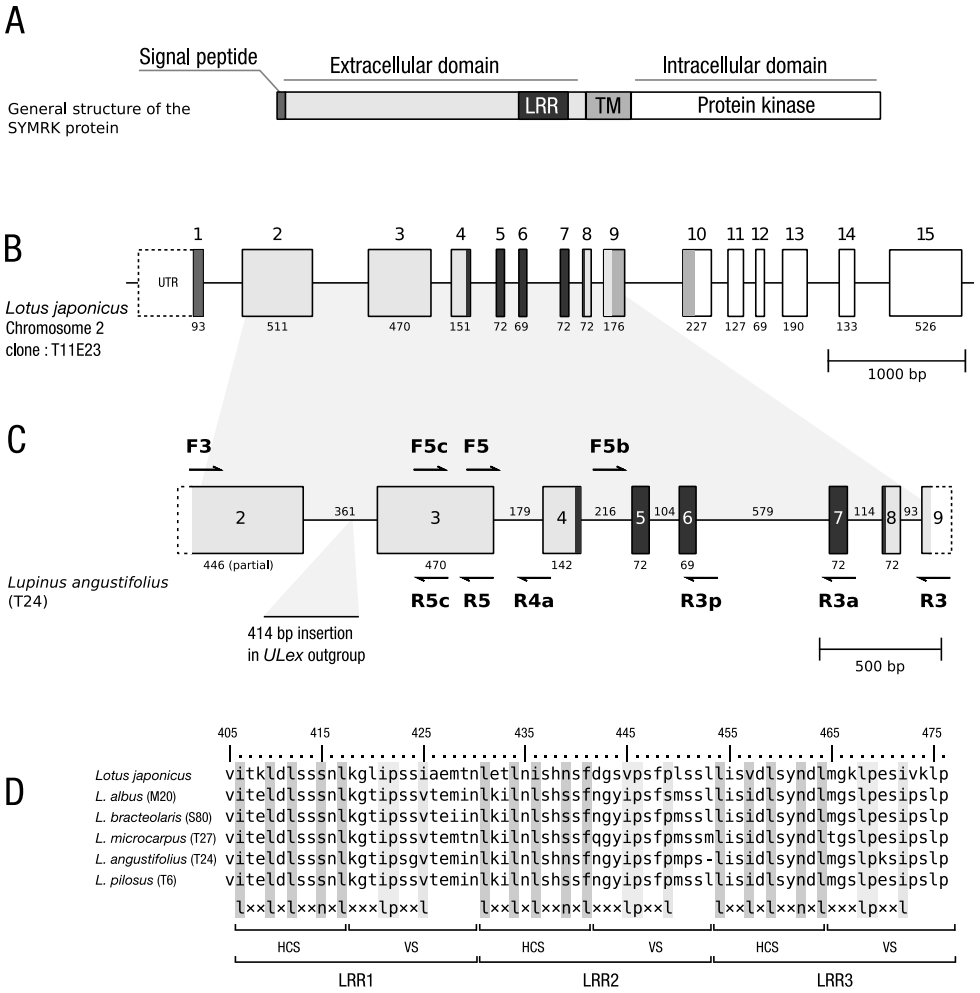

FIG. 6.4— A) General structure of the unfolded protein SYMRK: the extracellular domain contains three leucine-rich repeats (LRR), and is separated from the intracellular kinase domain by an hydrophobic region (TM). B) intronic/exonic structure of the *SymRK* gene observed in *Lotus japonicus*, and C) structure observed in *Lupinus angustifolius*, primers used for amplification are indicated by bold arrows. D) Comparison of predicted LRR amino acid sequences from five lupines. Each repeat is composed of two segments: a highly conserved segment (HCS) and a variable segment (VS).

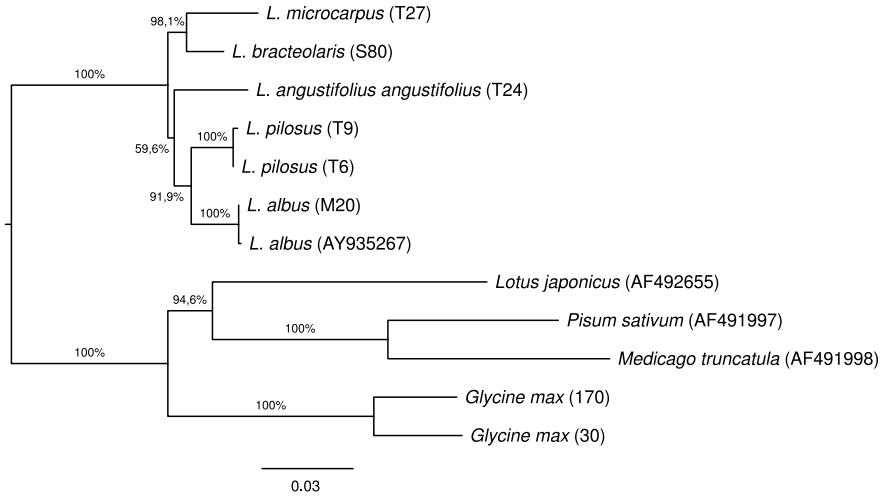

FIG. 6.5— Phylogenetic tree of predicted *SymRK* coding sequences (exon 2 to 8) based on M-Coffee alignment and maximum likelihood analysis using PHYML. Numbers above the branches represent the percentages of 1,000 bootstrap replicates. Sequences with a GenBank accession number are mRNAs, while sequence from *Glycine max* were predicted from the soybean genome sequencing project (scaffold 30 and scaffold 170, `http://www.phytozome.net`).

TAB. 6–7— Comparison of exonic-intronic structure, size and sequence divergence of the *SymRK* extracellular domain between *Lupinus* and *Lotus japonicus*.

Taxa	***Lotus***	***Lupinus***	**Mean divergence (%)**
exon 2	511	446	17.48
intron 2	391	353-361 (358)	30.25
exon 3	470	470	17.98
intron 3	144	119-186 (168.2)	39.51
exon 4	151	142	19.24
intron 4	185	174-216 (206.2)	32.65
exon 5	72	72	16.67
intron 5	104	103-104 (103.8)	21.56
exon 6	69	69	21.15
intron 6	230	538-579 (560.6)	29.39
exon 7	72	72	20.14
intron 7	101	94.114 (109.6)	26.64
exon 8	72	72	18.11
intron 8	85	90-93 (91)	36.91

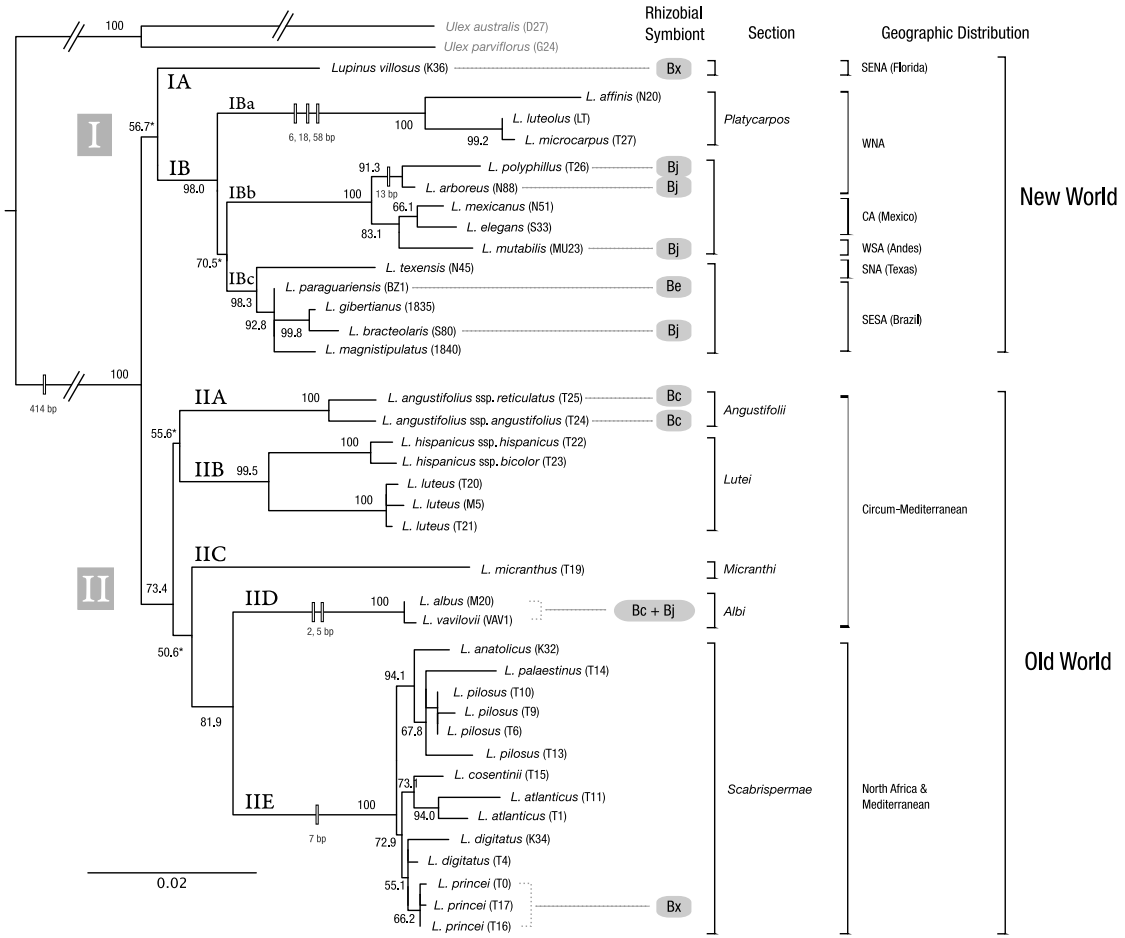

FIG. 6.6 — Phylogenetic analysis of *SymRK* sequences using the maximum likelihood method (1,000 bootstraps, values indicated in percents). Informative indels are indicated by empty rectangles along branches. Rhizobial symbionts are indicated in gray boxes: Bj, *Bradyrhizobium japonicum*; Be, *B. elkanii*; Bc, *B. canariense*; Bx, unknown *Bradyrhizobium*.

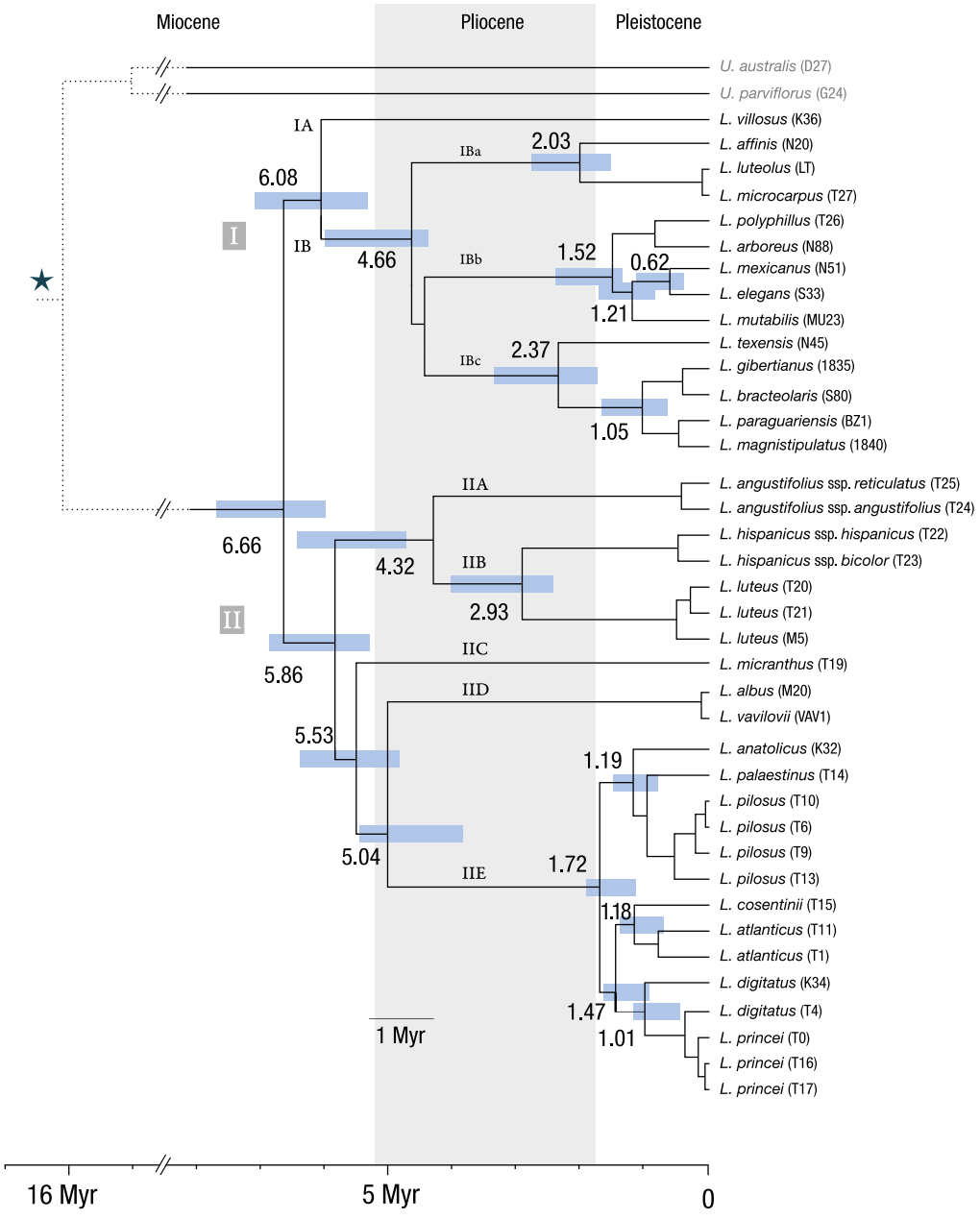

FIG. 6.7 — Phylogenetic analysis of *SymRK* sequences using a Bayesian analysis with the GTR+Γ model. Node ages are indicated along branches (in million years). Blue bars indicate the 95% confidence interval and calibration was made according to Hughes & Eastwood (2006) results.

B

G24	K36	S33	N88	BZ1	S80	1835	N45	T27	LT	N20	T25	T24	T22	T23	T20	T14	T10	T6	T1	K34	T4	T17	M20	T19	
0.54**	0.44**	0.46**	0.42**	0.51**	0.50**	0.49**	0.43**	0.49**	0.50**	0.48**	0.45**	0.43**	0.45**	0.46**	0.45**	0.54**	0.51**	0.51**	0.56**	0.52**	0.51**	0.51**	0.46**	0.56**	Lotus
	0.40**	0.38**	0.33**	0.37**	0.37**	0.36**	0.35**	0.47**	0.47**	0.48**	0.31**	0.32**	0.34**	0.34**	0.34**	0.45**	0.40**	0.40**	0.53**	0.41**	0.39**	0.39**	0.33**	0.53**	G24
		0.34**	0.25**	0.37**	0.41**	0.35**	0.24**	0.46**	0.47*	0.45*	0.43**	0.43**	0.44**	0.45**	0.36**	0.83	0.71	0.71	0.72	0.64	0.69	0.66	0.48*	0.94	K36
			0.31**	0.33**	0.39**	0.38**	0.24**	0.49**	0.50**	0.53*	0.36**	0.36**	0.37**	0.37**	0.32**	0.46**	0.42**	0.42**	0.53**	0.42**	0.41**	0.40**	0.35**	0.60*	S33
				0.20**	0.27**	0.24**	0.16**	0.39**	0.40**	0.35**	0.28**	0.28**	0.29**	0.29**	0.24**	0.36**	0.33**	0.33**	0.42**	0.32**	0.31**	0.31**	0.27**	0.49**	N88
					0.79	0.46	0.18**	0.55	0.57	0.53*	0.34**	0.34**	0.29**	0.29**	0.25**	0.49*	0.44**	0.44**	0.57**	0.42**	0.41**	0.41**	0.31**	0.68	BZ1
						0.57	0.21**	0.58	0.57	0.53*	0.33**	0.33**	0.29**	0.30**	0.26**	0.51*	0.43**	0.43**	0.58**	0.41**	0.40**	0.39**	0.31**	0.63	S80
							0.15**	0.6	0.63	0.57	0.33**	0.34**	0.28**	0.28**	0.24**	0.52*	0.44**	0.44**	0.60*	0.41**	0.41**	0.39**	0.28**	0.68	1835
								0.37**	0.38**	0.33**	0.26**	0.27**	0.25**	0.25**	0.23**	0.41**	0.35**	0.35**	0.55**	0.34**	0.33**	0.32**	0.25**	0.46**	N45
									41.82	0.34*	0.50*	0.52*	0.47**	0.47**	0.41**	0.67	0.64	0.64	0.66*	0.53**	0.59	0.59*	0.50**	0.82	T27
										0.37*	0.51*	0.53*	0.48**	0.48**	0.42**	0.68	0.65	0.65	0.67*	0.54**	0.6	0.6	0.50**	0.83	LT
											0.50*	0.48**	0.45**	0.45**	0.40**	0.67	0.62	0.62	0.61*	0.52**	0.6	0.59	0.48**	0.84	N20
												0.17*	0.34**	0.32**	0.26**	0.49*	0.40**	0.40**	0.51**	0.41**	0.38**	0.36**	0.30**	0.87	T25
													0.36**	0.35**	0.27**	0.52*	0.44**	0.44**	0.48**	0.42**	0.39**	0.38**	0.32**	0.76	T24
														0.89	0.19**	0.33**	0.27**	0.27**	0.50**	0.31**	0.24**	0.23**	0.21**	0.63*	T22
															0.19**	0.37**	0.31**	0.31**	0.53**	0.34**	0.28**	0.27**	0.25**	0.61*	T23
																0.43**	0.34**	0.34**	0.56**	0.37**	0.31**	0.31**	0.27**	0.61*	T20
																	1.33	1.33	0.65	0.56	0.65	0.64	0.33**	1.01	T14
																		NA	0.72	0.56	0.43	0.4	0.26**	0.92	T10
																			0.72	0.56	0.43	0.4	0.26**	0.92	T6
																				0.57*	0.68	0.67	0.45**	0.86	T1
																					0.51	0.5	0.35**	0.73	K34
																						0.00*	0.21**	0.91	T4
																							0.20**	0.87	T17
																								0.75	M20

A

Medicago	Pisum	Glycine	M20	S80	T24	T27	T6	
0.37**	0.35**	0.49**	0.44**	0.47**	0.42**	0.47**	0.49**	Lotus
	0.47**	0.48**	0.50**	0.48**	0.45**	0.49**	0.52**	Medicago
		0.51**	0.39**	0.40**	0.36**	0.40**	0.41**	Pisum
			0.42**	0.50**	0.45**	0.52**	0.46**	Glycine
				0.30**	0.37**	0.49**	0.27**	M20
					0.37**	0.58	0.42**	S80
						0.54**	0.50*	T24
							0.63	T27

FIG. 6.8 — Matrix of K_A/K_S ratios estimated between pairs of sequences of *SymRK* extracellular domain. Each ratio is a weighted average of 19 different models of molecular evolution tested by KAKS_CALCULATOR. Statistical significance (p-value) of each ratio is indicated in tables A and B as follows: $** < 0.01$; $* < 0.05$; shaded in light gray when p was higher than 0.05. **A**: pairwise comparison of sequences (including introns 2 to 8) of five lupine species with homologous *SymRK* mRNAs issued from four selected model legumes. **B**: pairwise comparison of a broader selection of lupines (22 samples) and one outgroup, based on a region including exons 2 to 6. Lupine samples are designated by their reference numbers following Table 6–5.

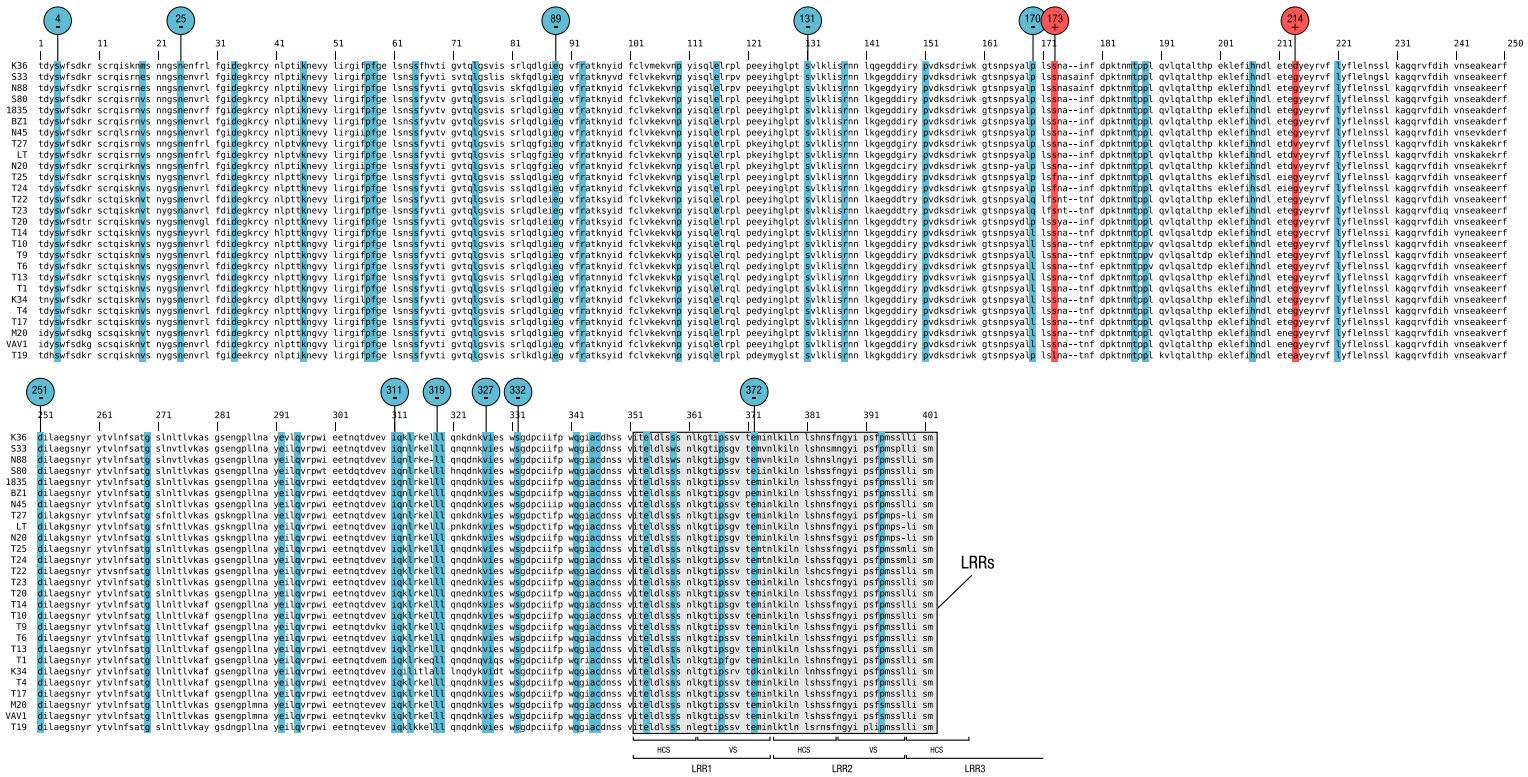

FIG. 6.9 — Matrix of predicted amino acid sequences from 27 *Lupinus* taxa (exons 2 to 6). The two positively selected positions are topped by red balloons. The 42 negatively selected positions are highlighted in blue. As for positively selected positions, positions reported by at least two of the three methods used are topped by blue balloons. Leucine rich repeat (LRR) region is indicated by a gray background.

6.4 Phylogénies combinées des régions ITS, ETS et *SymRK*

Dans cette partie, nous avons voulu combiner les données ITS, ETS et *SymRK*, selon le principe de *total evidence*, pour réunir le maximum d'information et améliorer notre connaissance des relations entre les grands groupes de lupins. Pour cette analyse, les taxons aux statuts ambigus *L. mariae-josephi* (MJ1) et *L. pilosus tassilicus* (A641), non-inclus dans l'article *SymRK*, ont été introduits ici. Considérant le caractère particulièrement énigmatique de *L. mariae-josephi*, une espèce récemment décrite en Espagne (Pascual, 2004), nous lui avons accordé un intérêt particulier en vue de mieux l'identifier et d'élucider sa position au sein du genre. Le travail sur cette espèce a fait l'objet d'un article spécifique, soumis à la revue *Genetic Resources and Crop Evolution*. Cet article est présenté page 169.

La recherche par JMODELTEST du modèle le plus adapté indique une dominance du modèle TIM3+Γ ($\ell = 12\,282,9725$) avec un poids de plus de 48 % ($w = 0,4856$; voir Tab. 6–8). Le paramètre $\alpha(\Gamma) = 0,7440$ indique une distribution des taux de variations se rapprochant d'une distribution normale. Le paramètre « positions invariantes » n'influence pas l'analyse ($p_{inv}(\mathrm{I}) = 0,0000$). La topologie moyenne pondérée, basée sur la totalité des modèles, indique que tous les clades reçoivent un support de 1. Ils sont donc stables et faiblement influencés par le choix du modèle. Le résultat de l'analyse phylogénétique et les valeurs de *bootstrap* sont présentés dans la Fig. 6.12 page 131.

TAB. 6–8 — Sélection de modèles pour la matrice ITS + ETS+*SymRK*. Sur les 88 modèles testés, 10 ont un $\Delta < 10$ et les 7 premiers représentent plus de 95 % du poids des différents modèles.

Modèle	ℓ	K	AIC_c	Δ	poids	poids cumulé
TIM3+Γ	12 282,9725	75	24 720,1270	0,0000	0,4856	0,4856
TIM3+I + Γ	12 282,9730	76	24 722,2410	2,1141	0,1687	0,6543
GTR+Γ	12 282,1374	77	24 722,6845	2,5575	0,1352	0,7894
TrN+Γ	12 286,1984	74	24 724,4672	4,3403	0,0554	0,8449
GTR+I + Γ	12 282,1385	78	24 724,8028	4,6758	0,0469	0,8917
TIM2+Γ	12 285,3696	75	24 724,9211	4,7941	0,0442	0,9359
TIM1+Γ	12 286,0860	75	24 726,3539	6,2269	0,0216	0,9575
TrN+I + Γ	12 286,1992	75	24 726,5803	6,4533	0,0193	0,9768
TIM2+I + Γ	12 285,3699	76	24 727,0348	6,9078	0,0154	0,9921
TIM1+I + Γ	12 286,0865	76	24 728,4680	8,3410	0,0075	0,9996

La combinaison des données ITS, ETS et *SymRK* produit une topologie proche de celle obtenue avec le gène *SymRK* seul. À l'exception de certaines branches terminales et du clade « Ancien Monde », tous les clades reçoivent un soutien fort. Le support du clade « Nouveau Monde » est nettement amélioré (56,7 à 84,9 %) tandis que celui de l'« Ancien Monde » diminue (73,4 à 58 %). Le soutien du clade *Angustifolii* + *Lutei* augmente considérablement (55,6 à 98,1 %), ainsi que celui du clade *Micranthi* + *Albi* + *Scabrispermae* (50,6 à 79,8 %).

La position de *L. villosus*, représentant des unifoliolés du Sud-Est de l'Amérique du Nord, en groupe-frère de l'ensemble des autres clades du Nouveau Monde est confirmée. La section *Platycarpos* (IBa) forme une lignée-sœur de l'ensemble des autres lupins ouest-américains (IBb), tandis que les lupins unifoliolés d'Amérique du Sud et leurs alliés du Sud des États-Unis (IBc) se placent en groupe-frère de cet ensemble IBa + IBb.

Dans l'Ancien Monde, deux clades principaux se dégagent. Un premier clade regroupant les sections *Angustifolii* + *Lutei*, et un deuxième clade regroupant les sections *Micranthi*, *Albi* et *Scabrispermae*. L'espèce *L. mariae-josephi*, introduite ici pour la première fois, vient se placer aux côtés de *L. micranthus* (86,5 %). Le taxon *L. pilosus tassilicus*, comme indiqué par les ITS, n'est pas lié à *L. pilosus* mais se place dans un deuxième sous-clade des *Scabrispermae* regroupant *L. digitatus*, *L. princei*, *L. atlanticus* et *L. cosentinii*.

Afin d'estimer l'âge du dernier ancêtre commun des clades décrits plus haut, la matrice ITS + ETS + *SymRK* a été soumise à une analyse bayésienne (méthode décrite page 83). Les cinq analyses indépendantes menées en parallèle convergent toutes vers une solution unique, comme indiqué dans la Fig. 6.10. La phylogénie obtenue est présentée dans la Fig. 6.11 page suivante. Les nœuds internes de l'arbre ont été datés en utilisant pour la racine l'âge de 16,01 millions d'années estimée par Lavin *et al.* (2005) pour la divergence entre *Lupinus* et les autres génistées. Les âges obtenus sont en moyenne 30 à 45 % plus grands que ceux estimés par l'analyse basée uniquement sur le gène *SymRK* (voir Fig. 6.7 page 125), sans toutefois que les résultats soient incompatibles, les intervalles de confiance se chevauchant.

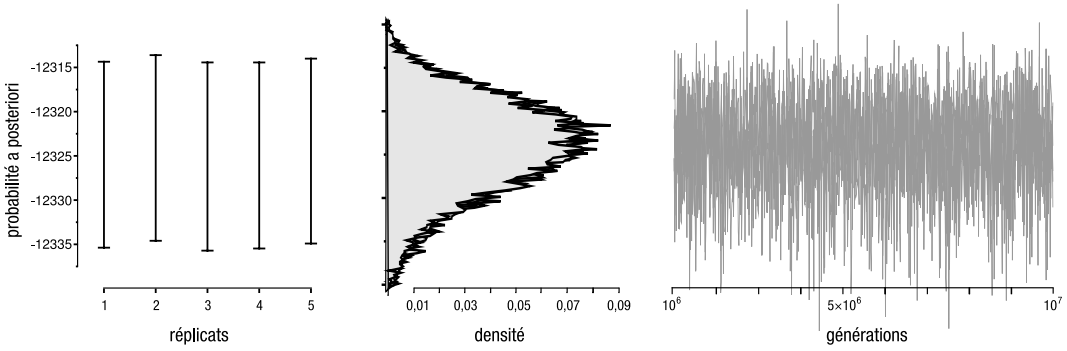

FIG. 6.10 — Analyse bayésienne de la matrice ITS + ETS + *SymRK* : vérification statistique. Les cinq réplicats convergent vers le même intervalle de valeur. La distribution des probabilités *a posteriori* suit une distribution normale et les cinq chaînes sont stabilisées dès le premier million de générations. Données produites par BEAST et analysées par TRACER.

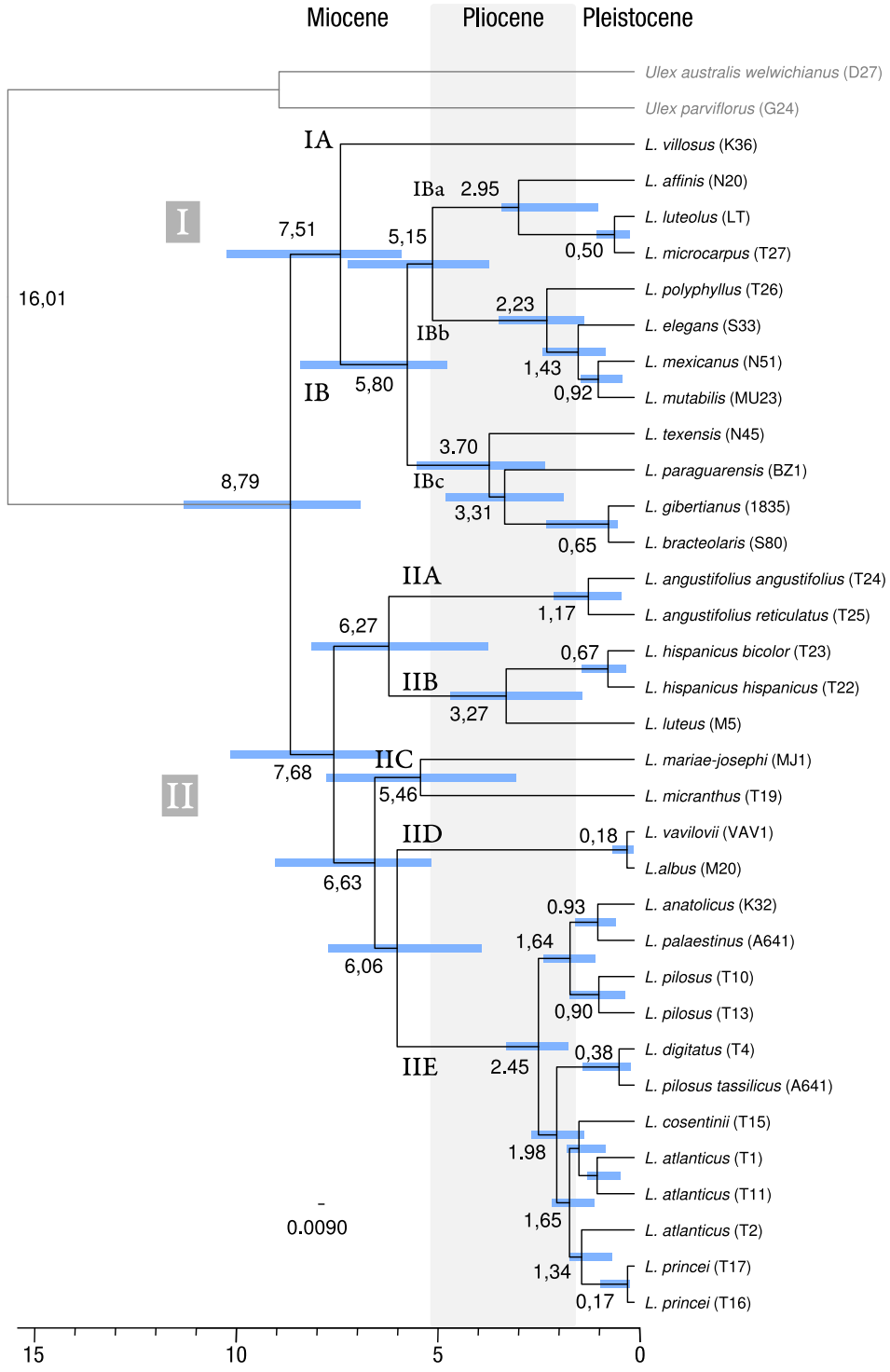

FIG. 6.11 — Phylogénie obtenue par analyse bayésienne de la matrice ITS + ETS + *SymRK* et calibrée selon les résultats de Lavin *et al.* (2005). Les dates sont indiquées en millions d'années et l'intervalle de confiance à 95 % concernant chaque âge est indiqué par une barre bleue.

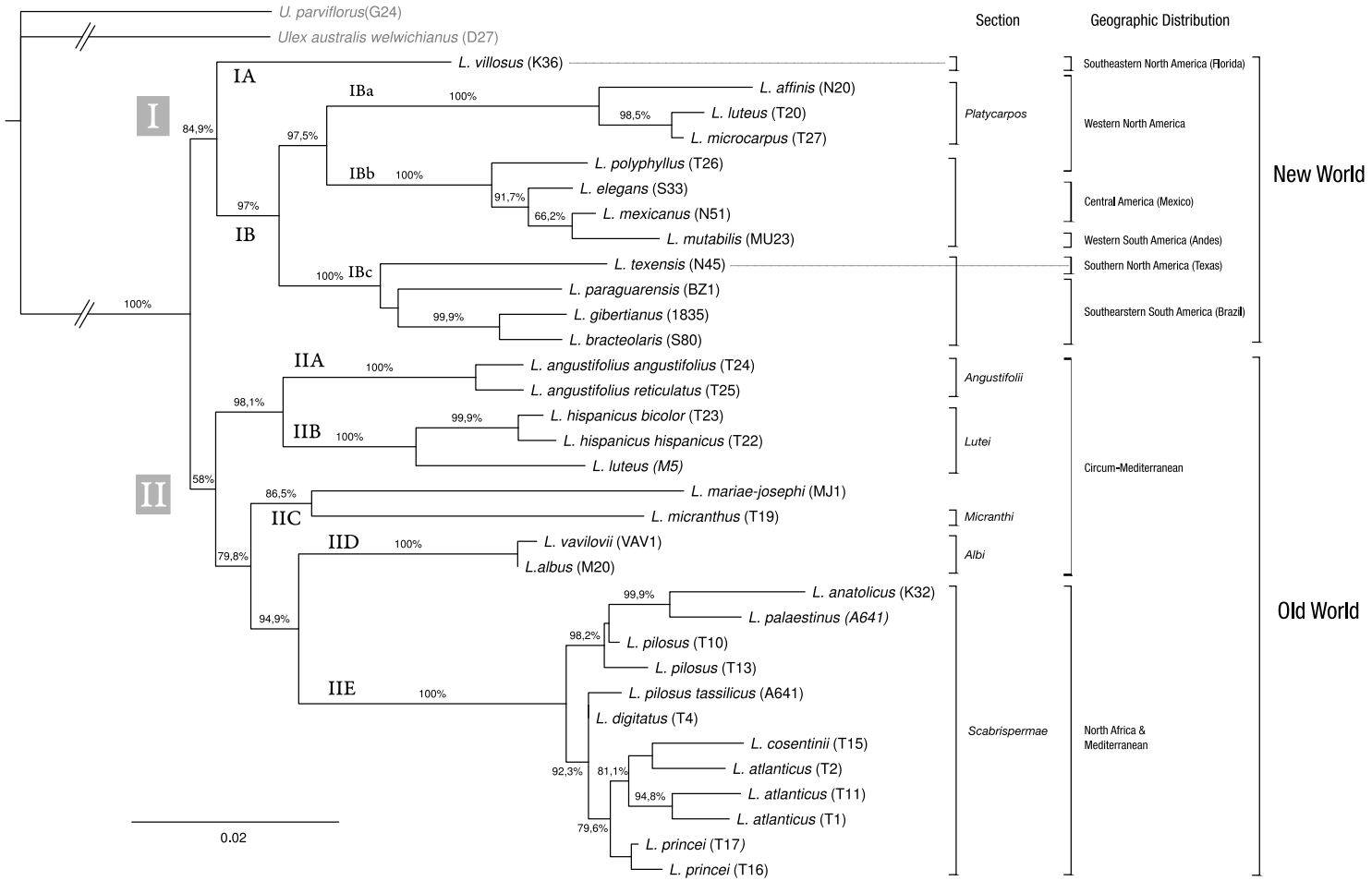

FIG. 6.12 — Phylogénie combinée des régions ITS, ETS et *SymRK*. Les analyses ont été réalisées selon la méthode du maximum de vraisemblance et le support des clades a été évalué par la méthode du *bootstrap* (1 000 réplicats, valeurs indiquées en pourcentage).

6.5 Conclusion

Dans ce chapitre nous avons utilisé trois jeux de données (ITS et ETS, deux espaceurs transcrits des gènes répétés ARNr et le gène *SymRK*) dans le but d'améliorer la phylogénie du genre *Lupinus*. Pour ce qui concerne les ITS, de nouvelles séquences sont venues compléter le jeu de données existant et ont permis de confirmer la position de certains taxons problématiques, comme *L. vavilovii* ou *L. pilosus tassilicus*. Le séquençage des ETS, utilisées ici pour la première fois chez *Lupinus*, a permis d'apporter de nouveaux caractères phylogénétiquement informatifs. Toutefois, ces caractères ont pour l'essentiel confirmés les clades déjà connus sans permettre de résoudre les relations entre les grands groupes de lupins.

L'étude du *SymRK* a permis de mettre en évidence deux caractéristiques importantes de ce gène : 1) il semble être en copie unique, et 2) toutes les séquences *SymRK* disponibles pour les fabacées semblent orthologues, c'est-à-dire descendant toutes d´une même séquence présente chez le dernier ancêtre commun des fabacées. L'analyse phylogénétique des séquences du *SymRK* à permis à la fois de conforter les clades déjà observés auparavant (Käss & Wink, 1997a ; Aïnouche *et al.*, 2004), mais a permis également de clarifier les relations entre les grands groupes. Par exemple, les lupins du Nouveau Monde forment un groupe monophylétique, à l'exception des lupins unifoliolés de Floride dont la position reste à confirmer. Dans l'Ancien Monde, le *SymRK* a permis de clarifier les relations entre *L. albus* et les *Scabrispermae*, et entre *L. angustifolius*, *L. luteus* et *L. hispanicus*.

La combinaison des jeux de données ITS, ETS et *SymRK* nous a permis d'aboutir à une phylogénie plus robuste, où les relations sont mieux résolues avec encore toutefois des incertitudes dans les relations entre les lignées à la base du genre. Si nos résultats sont d'une façon générale en accord avec les phylogénies récentes établies à partir d'autres jeux de données (Ree *et al.*, 2004 ; Hughes & Eastwood, 2006 ; Ch. S. Drummond, 2008), nous avons toutefois mis en évidence une incongruence significative concernant la position du groupe singulier des lupins unifoliolés de Floride. En effet, ceux-ci se rattachent aux lupins de l'Ancien Monde lorsque le gène LEGCYC1A est utilisé (Hughes & Eastwood, 2006), alors que cette relation n'est confortée ni par les ETS ni par le *SymRK* qui les rattachent aux lupins du Nouveau Monde.

Selon Lavin *et al.* (2005), le genre *Lupinus* aurait divergé du reste des génistées il y a environ 16 ± 4 millions d'années. En utilisant cet âge pour calibrer nos analyses phylogénétiques, nous avons pu dater l'apparition des différents clades de lupins. Le dernier ancêtre commun des lupins actuelles aurait vécu il y a environ 6 à 11 millions d'années. En 1974, Gladstones a émis l'hypothèse, reprise par Hughes & Eastwood (2006), que les lupins se seraient diversifiés à partir de l'Ancien Monde. En tenant compte des incertitudes sur les relations à la base du genre et en mettant de côté le cas des lupins de Floride, il est possible d'ébaucher le scénario suivant pour les lupins du Nouveau Monde. Un évènement de dispersion à longue distance aurait eu lien entre l'Ancien et le Nouveau Monde il y a 4 à 7 millions d'années. À partir de cet évènement, deux lignées de lupins se seraient diversifiées.

1. Une première lignée aurait donné naissance il y a 2 à 5 millions d'années aux

lupins occupant actuellement le Sud des États-Unis, qui auraient eux-mêmes essaimé vers le Sud-Est de l'Amérique du Sud pour donner naissance à un clade de lupins pluri- et unifoliolés. Ce résultat contredit l'hypothèse d'une dispersion du Sud vers le Nord des lupins unifoliolés émise par Dunn (1971).

2. Une deuxième lignée regroupe l'ensemble des lupins ouest-américains et se divise en deux sous-lignées : les *Platycarpos* (1 à 3 millions d'années) et une lignée non-nommée (1,5 à 3 millions d'années). Celle-ci aurait donné naissance à l'ensemble des lupins de l'Ouest de l'Amérique du Nord puis aux lupins d'Amérique centrale d'où aurait émergé il y a 0,5 à 1,5 millions d'années la branche des lupins andins. Ces derniers représentent aujourd'hui un des exemples les plus remarquables de radiation à une échelle continentale (Hughes & Eastwood, 2006).

Les lupins de l'Ancien Monde se sont séparés il y a 6 à 10 millions d'années en deux clades : *Angustifolii* + *Lutei* d'un côté et *Micranthi* + *Albi* + *Scabrispermae* de l'autre. Ces derniers, dits lupins à graines rugueuses, seraient donc dérivés de lupins à graines lisses (*Albi* et *Micranthi*) de la région méditerranéenne. Si cette divergence est ancienne, 5 à 9 millions d'années, le dernier ancêtre commun des *Scabrispermae* actuelles est plus récent (1,5 à 3 millions d'années). À partir de cet ancêtre, les *Scabrispermae* se sont séparées en deux sous-clades : le groupe *Pilosi* présent dans l'Est du bassin Méditerranéen (Grèce, Anatolie, Proche-Orient) et le groupe *L. cosentinii* + *L. atlanticus* + *L. digitatus* + *L. princei* présent dans les régions arides d'Afrique du Nord (Maroc, Mauritanie, Sahara central, vallée du Nil) à l'exception de *L. princei*, endémique du Kenya (région équatoriale). Cette dernière espèce aurait dérivé directement de *L. digitatus* (Égypte) il y a moins d'1,5 millions d'années.

À partir des données de cette phylogénie combinée et des données de la littérature, nous disposons maintenant d'un cadre phylogénétique relativement bien résolu, qui permet de mieux comprendre l'évolution du genre *Lupinus* et sa répartition géographique.

7

Diversité des rétrotransposons et variations de la taille des génomes dans le genre *Lupinus*

LA DIVERSIFICATION des angiospermes s'est accompagnée d'une variation remarquable de la taille des génomes. Cette variation n'est pas corrélée à la complexité des organismes ce qui soulève la question de la nature et de l'impact d'une telle variation, en terme évolutif et adaptatif. Outre la polyploïdie, l'accumulation d'éléments transposables peut faire varier de façon considérable la taille des génomes. Ces éléments peuvent donc jouer un rôle majeur dans la dynamique évolutive structurale et fonctionnelle des génomes de plantes.

Avec des nombres chromosomiques ($2n = 32$ à 52) et des quantités d'ADN variables (0,9 à 2,5 picogrammes par noyau), les lupins constituent un bon modèle pour examiner l'impact des éléments transposables sur la variation des génomes et sur l'adaptation et la diversification des plantes. Une attention particulière est accordée aux lupins occupant des habitats très contrastés, ayant le même nombre de chromosomes, mais présentant des différences de taille de génome remarquables : *L. micranthus*/*L. luteus* et *L. albus*/*L. Luteus* (pourtour méditerranéen) ; et *L. princei*/*L. atlanticus* (Afrique).

Dans ce contexte, la diversité et l'évolution de rétrotransposons des familles Ty1/*copia* et Ty3/*gypsy* — souvent impliqués dans les variations de taille de génome — ont été explorées pour la première fois au sein du genre *Lupinus* (fabacées ; génistoidées), à partir de l'analyse des séquences de leur transcriptase inverse.

L'analyse phylogénétique des séquences de transcriptase inverse montre que certaines sous-familles de rétrotransposons apparaissent comme le résultat d'amplifications spécifiques des lignées méditerranéennes ou africaines. Des estimations de l'importance relative des séquences de transcriptase inverse par PCR semi-quantitative et par FISH révèlent que le lupin méditerranéen à gros génome (*L. luteus*) contient plus d'éléments de type *copia* et *gypsy* que ceux à petits génomes (*L. micranthus*, *L. albus*), tandis que c'est la situation inverse qui est observé chez les lupins africains. Ainsi, les résultats suggèrent que différentes classes d'éléments transposables, et différents mécanismes peuvent être impliqués dans la variation de la taille des génomes au sein du

genre *Lupinus*.

L'ensemble de ces résultats sera intégré dans un article en cours de préparation et présenté ci-dessous.

Exploring retrotransposon diversity and evolution in *Lupinus* species (Fabaceae) with varying genome size in the context of species diversification and adaptation to contrasted environmental conditions

Frédéric Mahé[1], Christophe Biteau[1], Émilie Robin[1], Rémy Pasquet[2], Olivier Catrice[3], Spencer Brown[3], Valérie Huteau[4], Olivier Coriton[4], Marie-Thérèse Misset[1], Abdelkader Aïnouche[1]

[1]UMR CNRS 6553 Ecobio, Université de Rennes-1, Campus scientifique de Baulieu Bât 14A, F-35042 Rennes cedex, France.
[2]ICIPE–IRD, PO Box 30772 Nairobi, Kenya.
[3]Institut des Sciences du Végétal, UPR-CNRS 2355, F-91198 Gif-sur-Yvette cedex, France.
[4] UMR 118 INRA-AgroCampus Rennes INRA Centre de Rennes BP 35327 F-35653 Le Rheu Cedex, France.

Transposable elements, and particularly retrotransposons may account for much of the structural evolutionary dynamics of the plant genomes, and may provide a new potential for adaptive response to environmental changes and for species diversification. In this paper we explored the impact of retrotransposons on genome size evolution in *Lupinus* (Fabaceae) with an emphasis on Old World lupines occurring in very contrasted habitats, with the same chromosome number, but displaying tremendous genome size differences. Reverse transcriptase (*rt*) sequences from class I Ty1-*copia* and Ty3-*gypsy* retrotransposons have been isolated, cloned and sequenced for the first time in *Lupinus*. Phylogenetic analyses revealed a high diversity of *rt* sequences in *Lupinus* and Genistoid legumes, which show significant similarity levels (51-86%) with *rt* sequences of various LTR retrotransposon families known in other legume lineages. Although some subfamilies of retrotransposons are ubiquitous throughout the genistoids and obviously are of common and earlier origin, there are evidence indicating that some groups of retrotransposons have been specifically amplified in either the New World, the Mediterranean or the African lupine lineages. Additionnaly, copy number estimates of the *rt* sequences using semi-quantitative PCR-based method and molecular cytogenetic analyses (FISH) showed a more significant contribution of *copia*-like and *gypsy*-like elements to genome size differences among Mediterranean lupines than in genome size differences among African lupines. Thus, the results suggest different evolutionary patterns and mechanisms involved in genome size variation within the same genus.

Introduction

Genomes are dynamic structures, and genome size variation is a major component of that dynamics (X. Zhang & Wessler, 2004 ; Bennett & Leitch, 2005b ; Bennetzen *et al.*, 2005 ; Petrov & Wendel, 2006). In angiosperms, genome size range variation is very important, with a ratio superior to 2,000 between the smallest and the largest known genomes (Bennett *et al.*, 2000 ; Greilhuber *et al.*, 2006). Such a tremendous variability raises the question of the nature of the accumulated DNA and its impact on the fitness of organisms. Excluding polyploidy (whole genome duplication), the gain or loss of repeated DNA is the main cause of genome size variation (Bennetzen, 2002 ; Bennetzen *et al.*, 2005 ; Petrov & Wendel, 2006). Most

repeated DNA is made of transposable elements (TEs). Present in low-copy number in small genomes such as *Arabidopsis* (Terol *et al.*, 2001), TEs can represent more than 50% of large genomes such as *Hordeum vulgare* or *Zea mays* (R. B. Flavell *et al.*, 1977 ; Barakat *et al.*, 1997 ; SanMiguel & Bennetzen, 1998). Transposable elements are divided in two main classes, according to their transposition mode: with an RNA intermediate (class I) or not (class II). Class I elements mode of transposition allow them to rapidly increase in copy number (Bennetzen *et al.*, 2005 ; Capy, 2005). Therefore, their influence on genome size variation can be important (Vicient, Kalendar & Schulman, 2001 ; Vicient, Jääskeläinen *et al.*, 2001 ; Bennetzen, 2002 ; Vitte & Panaud, 2003 ; Bennetzen *et al.*, 2005 ; Neumann *et al.*, 2006 ; Piégu *et al.*, 2006).

Class I elements comprise two large families of retroelements, Ty1-*copia* and Ty3-*gypsy*, which has been found in all major lineages of green plants (Viridiplantae), including land plants and green algae (A. J. Flavell *et al.*, 1992 ; Voytas *et al.*, 1992 ; Kamm *et al.*, 1996 ; Friesen *et al.*, 2001 ; Derelle *et al.*, 2006). The ubiquity of retroelements has been interpreted as a footprint of their presence in genomes before the diversification of green plants (Gribbon *et al.*, 1999). Functional Ty1 and Ty3 elements encode a number of proteins, including a reverse transcriptase (*rt*) which plays a fundamental role in the amplification of the retroelements. Degenerate primers were designed in highly conserved regions to amplify a fragment of the *rt* for *copia* (Flavell et al., 1992) and *gypsy* (Friesen et al., 2001). Resulting phenetic analyses of the reverse transcriptases in various organisms revealed high degrees of sequence heterogeneity both within a single organism and among taxa (Ruas *et al.*, 2008, and references therein). Also, sequence analyses of the *rt*s, based on their degenerative rate, allowed identification of different groups of sequence similarity, suggesting that retroelements accumulate by successive bursts of activity separated by period of relative dormancy (Vitte et al., 2007).

The mechanisms triggering the activity of transposable elements appear to be stress-related (Wessler et al., 1995; Grandbastien, 1998; Waugh O'Neill et al., 1998; Quattrocchio et al., 1999; Kalendar et al., 2000; Liu & Wendel, 2000; Jiang et al., 2003; Pourtau et al., 2003; Tsukahara et al., 2009). Structural and functional variation resulting from the activity of transposable elements have important consequences on the genetic diversity, the adaptation and the evolution of organisms (McClintock, 1984; Capy et al., 1998; Bennetzen, 2002; Brandt et al., 2005; Biémont & Vieira, 2006).

In this paper, we explored the evolutionary patterns of Ty1-*copia* and Ty3-*gypsy* retrotransposons among *Lupinus* species (Fabaceae, Genisteae sensu lato) adapted to contrasted environmental conditions and exhibiting variable genome size.

Lupinus is a natural and diverse group of papilionoid legumes of about 300 species distributed in a wide range of eco-geographical conditions in both New and Old World. Over 90 % of the genus occurs in the New World, while only 13-14 annual species are native from the Mediterranean region and from North Equatorial Africa (Dunn, 1984; Planchuello-Ravelo, 1984; Gladstones, 1998; and references therein). Increasing efforts have been developed during the last decades to decipher the notorious complexity of the genus. Combined information from both traditional approaches and molecular phylogenetic investigations have significantly increased our knowledge on the evolutionary history of the genus, providing a valuable framework for understanding species diversification and interpretation of character evolution (Käss and Wink, 1997; Aïnouche and Bayer, 1999; Aïnouche et al., 2004; Citerne et al., 2003; Ree et al., 2004; Hughes and Eastwood, 2006; Drummond, 2008).

Lupines represent an aneuploid series exhibiting a remarkable variation of their genomes at both structural and DNA content levels (Gladstones, 1998; Conterato and Schifino-Wittmann, 2006; Naganowska et al. 2003, 2005). In the small and well known

group of Old World lupines, which is the focus in this study, chromosome numbers ranged from $2n = 32$ to 52 (Gladstones, 1998) and their genome sizes varied from $2C = 0.9$ to 2.5 pg (Naganowska et al. 2003). Interestingly, significant genome size differences were observed among species adapted to various eco-geographical conditions and with the same chromosome number, such as for instance: the Mediterranean species *L. micranthus* and *L. luteus* ($2n = 52$), or the ecologically divergent African species *L. atlanticus* and *L. princei* ($2n = 38$). It suggested a probable role of transposable elements, and most particularly of retrotransposons, often implicated in such variation. But no genomic resources neither data from previous punctual studies were available to test this hypothesis.

Thus the main purpose of this study was: (1) to evaluate the diversity of retroelements Ty1-*copia* and Ty3-*gypsy* in *Lupinus* species (and some other Genistoid representatives), and to estimate their contribution to genome size variation among species; and (2) to examine whether genome size differences can be correlated with lupine diversification and adaptation. Accordingly, special attention was directed towards Old World lupines with the same number of chromosomes, but displaying genome size differences.

Materials and Methods

Plant material and DNA extraction Fragments of the reverse transcriptase (*rt*) gene of Ty1-*copia* and Ty3-*gypsy* retroelements were obtained for 34 legume accessions, including 27 *Lupinus* species and 7 other Genistoid representatives (Table 7–1). Plants were grown in the greenhouse at the University of Rennes-1 (France). Total genomic DNA was extracted from 100 mg of fresh leaf material sampled from young plants, employing the Nucleospin Plant kit (Macherey-Nagel, Düren, Germany) and following manufacturer's instructions.

Flow cytometry In order to estimate the genome size of lupine taxa, fresh leaves were sampled and analyzed by flow cytometry at the Institut des Sciences du Vegetal (ISV–CNRS, Gif-sur-Yvette, France), using a Coulter EPICS Elite ESP flow cytometer and the propidium iodide staining method (Peña & Sánchez-Moreiras, 2001). *Solanum lycopersicum* ($2C = 1.90$ pg) and *Petunia hybrida* cv PxPc6 ($2C = 2.85$ pg) (Marie, 1993) were used as internal standards. Three independent measures were made for each sample and results with a coefficient of variation superior to 3% were discarded. Results are given for a diploid genome (2C) in picogramms (pg, see Dolezel, 2003; Greilhuber, 2005b).

Molecular phylogeny of lupines Species phylogeny included 27 samples corresponding to 12 Old World lupine species and subspecies, 2 representatives of the New World lupines, and 1 representative of the sister group of *Lupinus* (*Cytisus heterochrous*; Genisteae sensu stricto) to serve as outgroup (Table 7–1). Sequences of the internal and external transcribed spacers (ITS and ETS) of the nuclear ribosomal DNA repeats have been obtained following the procedures described in Aïnouche et al. (2004) and Chandler et al. (2001). Phylogenetic reconstruction was based on a maximum parsimony analysis of a data set combining both ITS+ETS regions, using PAUP* program version 4.0b10 (Swofford, 2003). Bootstrap method was used to estimate the robustness of the clades (Felsenstein, 1985b).

Amplification, cloning and sequencing of reverse transcriptase fragments reverse transcriptase sequences were amplified from Ty1-*copia* elements with primers U (5′-ACNGCNTTYYTNCAYGG-3′) and D (5′-ARCATRTCRTCNACRTA-3′) published by Flavell et al. (1992a), and from Ty3-*gypsy* elements using primers GyRT1 (5′-MRNATGTGYGTNGAY TAYMG-3′) and GyRT4 (5′-RCAYTTNSWNARYTTNGCR-3′) published by Friesen et al. (2001), following the procedure described in Alix and Heslop-Harrison (2004). PCR products were purified with a Macherey-Nagel Nucleospin Extract II kit and cloned with pGEM-T vector (Promega, Madison, WI). For each accession, 10 to 15 randomly chosen colonies were used

for plasmid DNA isolation with the NucleoSpin Plasmid kit (Macherey-Nagel). Sequencing was performed by Macrogen Ltd (Seoul, Korea), with the primer T7. Additional *rt* sequences from GenBank (Benson et al., 2009), representing retrotransposons from other Fabaceae (*Cicer*, *Glycine*, *Vicia*, *Vigna*) and Solanaceae (*Solanum*, *Lycopersicon*, *Nicotiana*), were included in this study for comparison (see Table 7–2 for accession numbers).

Analyses of *rt* sequence Raw sequences were treated by Lucy (Chou & Holmes, 2001; Li & Chou, 2004) in order to remove low-quality regions and vector sequences. The homology of each sequence with already known reverse transcriptase sequences has been verified by a two step validation against both nr nucleotides and nr proteins databases[1]. Sequences were first filtered by blastn with an e-value threshold inferior to 10^{-7} (Altschul et al., 1990). Sequences rejected in that first step were treated by blastx (e-value < 10^{-7}) in order to detect more distant identities. To avoid very large gaps and to facilitate the alignment process, remaining sequences were size-filtered and outlayers—sequences with less than 40% average identity with all other sequences—were eliminated using T-Coffee (Notredame et al., 2000). T-Coffee also permitted to measure the level of redundancy and to calculate pairwise-sequence similarities. Alignments of valid sequences were made with M-Coffee (Wallace et al. 2006; Moretti et al. 2007).

The PAUP program (Swofford, 2003) was used to perform a phylogenetic analysis based on the Neighbor-Joining (NJ) method (Saitou, 1987) with a nucleotide Kimura 2-parameters model. Support for groups was evaluated with 5,000 bootstrap replicates (Felsenstein, 1985b).

In order to evaluate the potential selective pressure exerted on the obtained *rt* sequences, we built a sub-set containing only sequences fully translatable (no stop-codons). In coding sequences, synonymous (Ka) and non synonymous (Ks) substitution rates ratios (Ka/Ks) are indicatives of the selective pressure exerted on a DNA region (Messier and Stewart, 1997; Yang and Belawski, 2000). Accordingly, Ka/Ks ratios close to 1 reflect neutral evolution, Ka/Ks ratios inferior to 1 indicate a negative or purifying selection, corresponding to high selective constraints, whereas Ka/Ks values greater than 1 are evidence for a positive or relaxed selection, suggesting adaptive evolution. Resulting data sets (*copia* and *gypsy*) were analyzed by KaKs_Calculator (Zhang et al., 2006), using maximum-likelihood methods and 19 different models of evolution. As stated by Zhang et al., model averaging can reduce biases arising from model selection. We therefore chose the model-averaged method.

Fluorescence in situ Hybridization Chromosomal distribution and abundance of reverse transcriptase were visualized by fluorescence in situ hybridization. Root tips of approximately 3 mm were sampled from germinated seeds and treated in the dark with 0.04% 8-hydroxiquinoline (2 h at room temperature and 2 h at 4°C), followed by 48 h at 4°C in 3:1 ethanol:acetic acid and a conservation at –20°C in 70% ethanol until use. After being washed in distilled water for 10 min, and treated 15 min by a 0.01 M pH 4.5 citrate solution, root tips were incubated at 37°C for 30-35 min in an enzymatic mixture (5% cellulase and 1% pectolyase solution). After removal of the enzymatic solution, meristematic root tips were subjected to hypotonic treatment by addition of ultra pure water. After 30 min, root tips were squashed on a slide in a drop of 3:1 ethanol:acetic acid, and air-dried before further use. Labelled *rt* sequences were obtained from targetted genomes by PCR amplification with the PCR DIG Labelling Mix, according to manufacturer's instructions (Roche, Basel, Switzerland). An hybridization mixture was prepared with 25 µl of 100% formamide, 10% of dextran sulfate (0.5 g/ml), 5 µl of 20×SSC, 0.6 µl of SDS (0.2 g/ml), 5 µl of labelled 45S rRNA (positive control), approximately 100 ng of labelled *rt* sequences, 10 µg of denatured salmon sperm DNA (100 mg/ml) and ultra pure water for

1. `http://www.ncbi.nlm.nih.gov`

a final volume of 50 µl. Before hybridization, slides were washed in 2×SSC and incubated at 37°C for 1 h. Slides were then rinsed out 3 min in 2×SSC at 42°C (step repeated twice), treated with 100 ml of pepsine at 100 µg/ml during 20 min at 37°C, rinsed out 3 min in 2×SSC at 42°C (step repeated twice) and plunged in a solution of paraformaldehyde (0.05 g/ml) and NaOH (0.01 M) during 10 min. Slides were rinsed out 3 min in 2×SSC (step repeated twice), plunged 2 min in 70% deionized formamide and dehydrated by 3 successive 3 min baths in ice-cold 70%, 90% and 100% ethanol. Slides were air-dried at room temperature during 2 h.

After a denaturation step (6 min at 92°C), the hybridization mixture was applied to the slides (50 µl/slide). Slides and hybridization mixture were placed in a humid chamber at 37°C for about 12-14 h. In order to detect the hybridization signal, 50 µl of a mixture containing 40 µl of anti-DIG-FITC at 200 µg/ml (Roche) and 340 µl of 5% bovine serum albumine were added to the slides. After 1h at 37°C, the slides were washed 5 min in 4×SSC-Tween at 42°C (step repeated 3 times). Dried slides were then stained by a drop of 4',6-diamidino-2-phenylindole (DAPI). Cells were observed with an Olympus BX-51 microscope at 1,000× and digitally documented with a Pixera Penguin 600ES CCD camera.

Semi-Quantitative PCR Semi-quantitative PCR was performed in order to estimate the ratios of *rt* copy numbers in targeted remarkable couple of species, *L. micranthus*/*L. luteus* and *L. princei*/*L. atlanticus*. Prior to the experiment, concentrations of DNA extractions (number of nuclei per unit of volume) were estimated and normalized. PCR process, identical to the protocol described above, was stopped every two cycles, from cycle 13 to 35, yielding a total of 12 samples for each taxa. Samples were compared by gel electrophoresis, and can be interpreted as a kinetic of amplification reflecting the number of *rt* copies initially present in the PCR reactions. In ideal conditions, the copy number N of the targeted sequence doubles every PCR cycle ($N = N_0 \times 2^n$). Consequently, the copy number ratio between two taxa a and b can be calculated from the number of cycles (n) it takes to reach the same level of amplification ($2^{n_a}/2^{n_b}$). The value obtained is the upper-most limit and the true ratio is probably lower.

Results

After the identification and filtering of obtained sequences, a total of 380 putative *rt* fragments were retained: 260 Ty1-*copia* (accession numbers GU189754 to GU190013) and 120 Ty3-*gypsy* (accession numbers GU190014 to GU190133).

Ty1-*copia* Individual *rt-copia* fragments ranged in size from 248 to 295 bp (mean 266.4, median 267.0) and 50% of fragments were 265-268 bp long (see Fig. 7.2a). The distribution of pairwise sequence similarities between the 260 fragments was trimodal, with to main peaks at 56% and 65% of similarity and a third lower peak at 90% of similarity (see Fig. 7.2b). Of the 260 sequences analyzed, 122 (47%) contained stop codons revealing an relatively important number (138 or 53%) of potentially functional *rt* genes. The logo[2] (see Fig. 7.2c) generated from the alignment of these potentially functional sequences showed highly conserved positions, especially positions 45 to 54 and 93 to 98, the latter corresponding to the primer D.

Ty3-*gypsy* Individual *rt-gypsy* fragments ranged in size from 155 to 464 bp (mean 405.6, median 422.0) and 50% of fragments were 417-423 bp long (see Fig. 7.2d). The distribution of pairwise sequence similarities between the 120 fragments was trimodal, with to main peaks at 55% and 71% of similarity and a third lower peak at 91% of similarity (see Fig. 7.2e). Of the 120 sequences analyzed, 80 (66%) contained stop codons and 40 (33%) were potentially functional *rt* genes. The logo (see Fig. 7.2f) generated from the alignment of these potentially functional sequences showed some highly conserved positions scattered along the alignment, especially

2. `http://www.sciences.univ-nantes.fr/lina/bioserv/WebLogo/logo.cgi`

positions 5 to 9 corresponding to the primer GyRT1.

Analyses of Ka/Ks ratios Analyses of Ka/Ks ratios of coding *copia* and *gypsy* sequences showed an unexpected strong purifying selection. Values ranged from 0.00 to 1.35 with a mean and a median of 0.25, and were strongly supported by *p*-values. This results could indicate a slower than expected rate of accumulation of mutations in *rt* sequences post-transposition.

Phenetic analysis Combined analyses of the entire set of sequences (*rt-copia* and *rt-gypsy*) from all analyzed species revealed that certain types of sequences are present in genomes of all taxa studied. Ty1-*copia* sequences clearly segregate from Ty3-*gypsy* sequences (respectively left and right part of the Fig. 7.4a). Most *copia* fragments fell into three major clades (Fig. 7.4a, clades labeled 1, 2 and 3) besides a small clade (unlabelled) containing only nine distantly related elements from Mediterranean species. The three major *copia* clades contain sequences from all main lineages of lupines (see color legend on Fig. 7.4a). Clade *copia*-2 contains very divergent sequences (long branches) originating from all lupine lineages and Genistoids, suggesting the presence of these elements since the divergence of Genistoids. Clades *copia*-1 and *copia*-3 cluster very closely related and highly similar sequences (short branches) probably resulting from a recent amplification event. Among these groups of sequences showing high or weak identities, there is evidence from colored branches in Fig. 7.4a that several of them are related to specific lineages of lupins.

The same pattern is observed in the *gypsy* part of the tree, with two well-separated branches, *gypsy*-1 and *gypsy*-2. The clade *gypsy*-2 represents ubiquitous sequences, most likely deriving from an ancient retrotransposition event suggesting a recent lineage specific amplification. The branch *gypsy*-1 is divided in two sub-clades: the *a* clade contains a mixture of sequences from outgroups and Mediterranean lupine species, whereas the *b* clade clusters almost exclusively very closely related sequences from African lupines.

Sequences cloned from the different lineages of lupines or from outgroups are not randomly distributed in the tree. Sequences from Fabaceae and Genistoids are distributed in all major clades of the tree and are mainly grouped near the basis of each main clades (see for example groups 1, 2, 3 and 4 in Fig. 7.4b). Sequences from New World species also tend to group together (see groups 1, 2 and 3 in Fig. 7.4c) and are also an important component of the clade *copia*-2. Sequences from Mediterranean species are gathered in three clusters, each reflecting three levels of sequence similarity: the group 1 (see Fig. 7.4d) is composed of divergent sequences whereas groups 2 and 3 are composed of highly similar sequences. Among the latter, some groups are remarkable for being very specific to the Mediterranean lupines lineages displaying large genomes (*L. luteus*, *L. hispanicus*, and *L. angustifolius*) as indicated by arrow in Fig. 7.4d. Finally, sequences from *Scabrispermae*, or African species, are the ones with the most interesting distribution. Some sequences are scattered in clades *copia*-2, *gypsy*-2 and *copia*-3, but most of them are grouped in two large clusters of very similar sequences (see groups 1 and 2 in Fig. 7.4e). These two clusters indicate a recent burst of activity, for both Ty1-*copia* and Ty3-*gypsy* elements, specific to the branch of African lupines.

Semi-quantitative PCR In order to estimate the relative number of *rt* copies present in the genomes of two remarkable couples of lupine species, a semi-quantitative PCR technique was used. For the Mediterranean species, *L. micranthus* and *L. luteus*, we observed a difference of amplification, for both Ty1-*copia* and Ty3-*gypsy* *rt* sequences. For *copia*, the minimal level of detection is reached at different times: $t = 21$ for *L. luteus* and $t = 27$ for *L. micranthus* (see Fig. 7.6). This results clearly indicates the presence of more copies of *rt-copia* in *L. luteus* ($2C = 2.4$ pg) than in *L. micranthus* ($2C = 1.0$ pg). This difference can be capped at the theorical maxima of $2^6 = 64$ times more copies. The pattern

observed in African species is different. *Lupinus princei* ($2C = 1.0$ pg) reaches the level of detection earlier than *L. atlanticus* ($2C = 1.7$ pg), respectively $t = 17$ and $t = 23$. Surprisingly, the small genome appears to contain more copies of *rt-copia* ($max = 64$) than the large genome. For *gypsy* sequences, a similar trend is observed (results not shown).

Discussion

Genome size variation and *Lupinus* phylogeny Flow cytometry estimates of the 2C DNA content available from previous studies (Naganowska et al., 2003; and references therein), and confirmed for the samples surveyed here, showed a fairly remarkable variation of the genome size among the Old World lupines, ranging from 1.0 to 2.4 pg per nucleus. It is worth noting, that a similar range of variation (1.08 to 2.68 pg) was found among the New World species (Naganowska et al., 2005). When re-examined in a phylogenetic context, no unique pattern of genome size variation was apparent and different situations may be observed among the Old World species. The range of variation was greater among the smooth-seeded Mediterranean (1.0 to 2.4 pg). In the Mediterranean lupines, genome size is correlated with phylogenetic relationships, as seen for some couples of closely related species with similar chromosome numbers: *L. micranthus* ($2n = 52$; $2C = 1.07$ pg) *L. albus* ($2n = 50$; $2C = 1.16$ pg) or *L. luteus* ($2n = 52$; $2C = 2.42$ pg) *L. hispanicus* ($2n = 52$; $2C = 2.17$ pg). Interestingly, *luteus* and *hispanicus* are ecologically restricted to sandy soils, while *albus* and *micranthus* (with smaller genomes) have a wide edaphic range. In contrast, *L. angustifolius*, which displays a smaller chromosome number ($2n = 40$) and a significant genome size ($2C = 1.89$ pg), showed closer phylogenetic relationships to species with $2n = 52$ and bigger genomes (*L. luteus*-*L. hispanicus*). Previous molecular-based estimates indicated that *L. micranthus* and *L. albus* represent early diverged lines in the Old World (ca. 6.5-4.5 My ago) relative to the other Mediterranean lines (Hughes and Eastwood, 2006; Mahé et al., in prep.). One plausible hypothesis here is that genome size increase is a rather derived and lineage-specific process in the Mediterranean smooth-seeded lupines. However, testing this hypothesis implies to more accurately resolve some uncertainties regarding basal relationships among the Old World lupine. Genome size also varied significantly among the African rough-seeded lupines (ca. 1.0, 1.4 or 1.7 pg), but no obvious correlation can be established with their chromosome numbers. In contrast, the most remarkable observation is that the two species displaying $2n = 38$ chromosomes, which occur in very contrasted environmental conditions and exhibit a great genome size difference, *L. princei* (from Kenya; $2C = 1.0$ pg) and *L. atlanticus* (from Morocco; 1.7 pg), are phylogenetically closely related within a small group also including *L. digitatus* (from Egypt; $2n = 36$; $2C = 1.4$ pg). Thus there is evidence that the latter three species experienced consistent genomic changes, at both chromosomal and DNA content levels, after their divergence from a common ancestor, likely during the late Pleistocene (< 1 My, according to Mahé et al. in prep.). Moreover, the placement of *L. princei* within this group suggests that the small genome of this species might results from a rapid DNA loss following the colonization and adaptation to new tropical environmental conditions from a North African rough-seeded line. Besides, regarding the mean 2C value of about 1.4 pg reported for the rough-seeded lupines (Naganowska et al., 2003) which includes other species not examined here, it seems probable that the big genome size of *L. atlanticus* results from mechanisms of DNA accumulation in this lineage during its own evolutionary history in western North Africa.

Therefore, different patterns of genome size evolution (expansion/contraction) may be observed within the Old World lupines, which may or may not be easily correlated to the phylogeny. This was also mentioned for other taxa, and it was shown in genus *Hordeum* that DNA amount may vary more

rapidly than the process of speciation (Jakob et al., 2004). Accordingly, there are evidence demonstrating inter-crossing abilities among the recently diverged rough-seeded lupines (Carstairs et al., 1992), which is not the case among the older Mediterranean smooth-seeded ones (Plitmann and Pazy, 1984). Cumulated data and studies in plants, show that genome size expansion or contraction are not unidirectional, but are repeated and episodic events experienced by organisms along their evolutionary history, through various mechanisms of DNA accumulation vs. removal (Bennetzen, 2002; Devos et al., 2002; Petrov and Wendel, 2004; Hawkins et al., 2008, 2009). Re-examination of the data available for the Old World in a phylogenetic context is consistent with a bidirectional evolution of genome size with evidence of differential rates and timescale of variation among lineages. Differential accumulation vs. removal of repetive DNA, and particularly of retrotransposons, has been widely demonstrated as being one of the most important mechanisms responsible for genome size variation among closely related species (Wicker et al., 2001; Petrov and Wendel, 2004; Hill et al., 2005; Hawkins et al., 2006, 2009; Piégu et al., 2006).

Diversity and lineage-specific amplification of retrotransposons This study presents the first evaluation of the diversity of the Ty1-*copia* and Ty3-*gypsy* families of retrotransposons in *Lupinus* species and in some representatives of the Genistoid alliance. A remarkable diversity of Ty1-*copia* elements is detected within genomes of various genistoid legumes, based on the analysis of their transcriptase reverse domain (*rt*). The latter showed significant similarity levels with those of other papilionoid groups, such as *Cicer* (52,5 to 88,4%), *Vigna* (51 to 86%), *Vicia* and Soybean. Also, part of the genistoid *rt* sequences displayed a high similarity level with those representing a family of Ty1-*copia* (Tto1) described in tobacco (Solanaceae; Asteridae by Hirochika et al., 1996).

The scattered placement of extra-genistoid reverse transcriptase observed in the phylogeny of the lupine and genistoid sequences indicates that these *copia* and *gypsy* subfamilies are of ancient origin and common to legumes and Eudicots (Flavell et al., 1992; Matsuoka and Tsunewaki, 1996). Phylogenetic analysis of the *rt* sequences showed that *copia* and *gypsy* elements diversified into three and two main lineages, respectively. Within each of these lineages two main kinds of distinct monophyletic groups of sequences were identified: groups containing divergent sequences from various Genistoid taxa (including lupines), which represent elements likely resulting from ancient retrotranspositional events; and groups including weakly divergent sequences isolated from the same taxonomic lineage, indicating recent lineage-specific transpositional activities. Two main classes of lineage-specific groups may be distinguished: those deriving from short branches, indicating a recent origin from recently derived elements; and those deriving at the extremity of long branches, likely indicating a recent origin from an older but still active element. Thus, additional to evidence of ancient burst of activity which likely occurred before divergence of the Genistoids (for instance in the copia 2 lineage, Fig. 7.2), our data provided clear evidence of several recent lineage-specific amplifications of both *copia* and *gypsy*, specific to: the extra-lupine genistoids, to the New World lupines, or specific to either the Mediterranean or the African Old World lineages. Moreover, despite the limits of the sampling, the data allowed detection of some *copia* and *gypsy* groups that likely represent amplifications more specific to the Mediterranean species displaying big genome size (*L. luteus*, *L. hispanicus* and *L. angustifolius*). Therefore, additional to the general overview of the diversity of the retrotransposons families, examination of the *rt*-phylogeny also allowed detection of successive bursts of activities in *Lupinus*, including within specific lineages. This is in accordance with the groups of similarity revealed by the distribution of pairwise sequence identities among the 260 *copia* and 120 *gypsy* fragments analyzed here. These data identified three main transpositional events

(illustrated by 3 peaks in Fig. 7.2), which likely occurred at nearly the same time in the recent evolutionary history of the genus. Additionally, the data showed that retrotransposition events may vary in intensity according to the retrotransposon family (*copia* or *gypsy*), at different evolutionary times. For instance, there is evidence from present data that *copia* elements apparently experienced a much higher amplification (see peak2 in Fig. 7.2) than *gypsy* elements in the recent past of the genus. Such pattern of successive bursts of retrotransposons with differential lineage-specific amplification, including differential-retrotransposon amplification intensity, emerge more and more as a common and recurrent process experienced by plants during their evolutionary history, as this is well exemplified from recent studies in rice and cotton (Piégu et al., 2006; Hawkins et al., 2008, 2009).

Contribution of retrotransposons to genome size variation in Old World lupines Fluorescent in situ hybridization (Fig. 7.5) and preliminary comparative estimates of the *rt* sequence copy number using a semi-quantitative PCR-based method (Fig. 7.6) show a much higher accumulation of both *copia* and *gypsy* retro-transposons families in the large genome (*L. luteus*) than in the small one (*L. micranthus*), in the Mediterranean lupines.

The other situation found among the African lupines showed that, though *copia* and *gypsy* elements are present, they are not responsible of the great genome size difference observed between the closely related species *L. atlanticus* and *L. princei*, which suggests that DNA components other than *copia* and *gypsy*-element types (e.g. other repetitive sequences including other transposable element categories, satellite DNA, ...) are involved in genome size evolution affecting this lineage. Thus, different evolutionary patterns and mechanisms appear to have operated in genome size variation among different lupine lineages within the same genus. Further analyses involving large scale comparative sequencing of various representatives of the different lupine lineages are needed to elucidate this question.

References

(see p. 191)

TAB. 7–1 — List of *Lupinus* and outgroup taxa included in this study. Samples are presented with their origin, geographic distribution, and reference number. Abbreviations: OW, Old World; NW, New World; NA, North America; SA, South America; CA, Central America; Afr, Africa; Med, Mediterranean. USDA, US Department of Agriculture, Washington; INAE-DZ, Institut National d'Agronomie, El-Harrach, Algérie; AKA, Abdelkader Aïnouche; INRAL, INRA, Lusignan, France; BRA, EMBRAPA, Brasil. RP, Rémy Pasquier; UFRGS, Universidade Federal do Rio Grande do Sul; IV, Institut Vavilov, St Petersburg; WADA, Western Australia Department of Agriculture.

Taxon	**2*n***	**Origin/Distribution**	**Sample source & Reference number**
L. affinis	48	Oregon/NW, West NA	USDA/504315/N20
L. albus	50	Algeria/OW, Med	INAE-DZ/M20
L. anatolicus	42	Turkey/OW, Afr	AKA/K32
L. angustifolius ssp. *reticulatus*	40	France/OW, Med	AKA/T25
L. angustifolius ssp. *angustifolius*	40	Algeria/OW, Med	AKA-M1/T24
L. atlanticus	38	Morocco/OW, Afr	USDA/384612-FM83/T1
——	38	Morocco/OW, Afr	INRA-SAPF/T11
——	38	Morocco/OW, Afr	USDA/384613-FM87/T2
L. bracteolaris	32-34	Brazil/NW, South-East SA	USDA/404349/S80
L. concinnus	?	USA/NW	N19
L. cosentinii	32	?/OW, Med	INRAL-FR/T15
L. diffusus	?	Florida/NW	K35
L. digitatus	36	Egypt/OW, Afr-Med	WADA-PI26877/T4
L. elegans	48	Mexico/NW, West SA	USDA/185099/S33
L. hirsutissimus	?	USA/NW	AKA/N85
L. hispanicus ssp. *bicolor*	52	Spain/OW, Med	USDA/PI 384554/T23
L. hispanicus ssp. *hispanicus*	52	Portugal /OW, Med	USDA/384555/T22
L. luteus	52	Algeria/OW, Med	AKA/M5
——	52	Algeria/OW, Med	AKA/T20
——	52	Algeria/OW, Med	AKA/T21
L. mariae-josephi	52?	Spain/OW, Med	H. Pascual/MJ1
L. micranthus	52	Algeria/OW, Med	AKA/T19
——	52	Algeria/OW, Med	T 28
L. mutabilis	48	Perou/NW, West SA	INAE-DZ/S35/MU23
L. nanus	48	USA/NW	N42
L. palaestinus	42	Near-East/OW, Afr-Med	INRA-FR/T14
L. paraguariensis	36	Brazil/NW, East SA	BRA-02828/BZ1
L. pilosus	42	Algeria/OW, Afr-Med	INAE-DZ/T6
——	42	Algeria/OW, Afr-Med	INAE-DZ/T9
——	42	North-Africa/OW, Afr-Med	USDA/W6 PI 11995/T13
L. pilosus tassilicus	?	Lybia/OW, Afr	AKA/A641
L. polyphyllus	48	USA/NW, NA	USDA/504404/T26
L. princei	38	Kenya/OW, Afr	WADA P 23021/T0
——	38	Kenya/OW, Afr	RP Chyulu 1800/T16
——	38	Kenya/OW, Afr	RP Chyulu 1915/T17
L. texensis	36	USA/NW, South NA	USDA/577291/N45
Anarthrophyllum cumingii	?	?/NW, South SA	AKA/201
Argyrolobium uniflorum	?	OW	AKA/G25
Chamaecytisus mollis	?	OW	AKA/C84
Crotalaria podocarpa	?	OW	AKA/K50
Cytisus heterochrous	?	OW	AKA/G8
Genista tinctoria	?	OW	AKA/G56
Thermopsis rhombifolia	?	NW	AKA/G46
Ulex parviflorus	?	Spain/OW, Med	AKA/G24

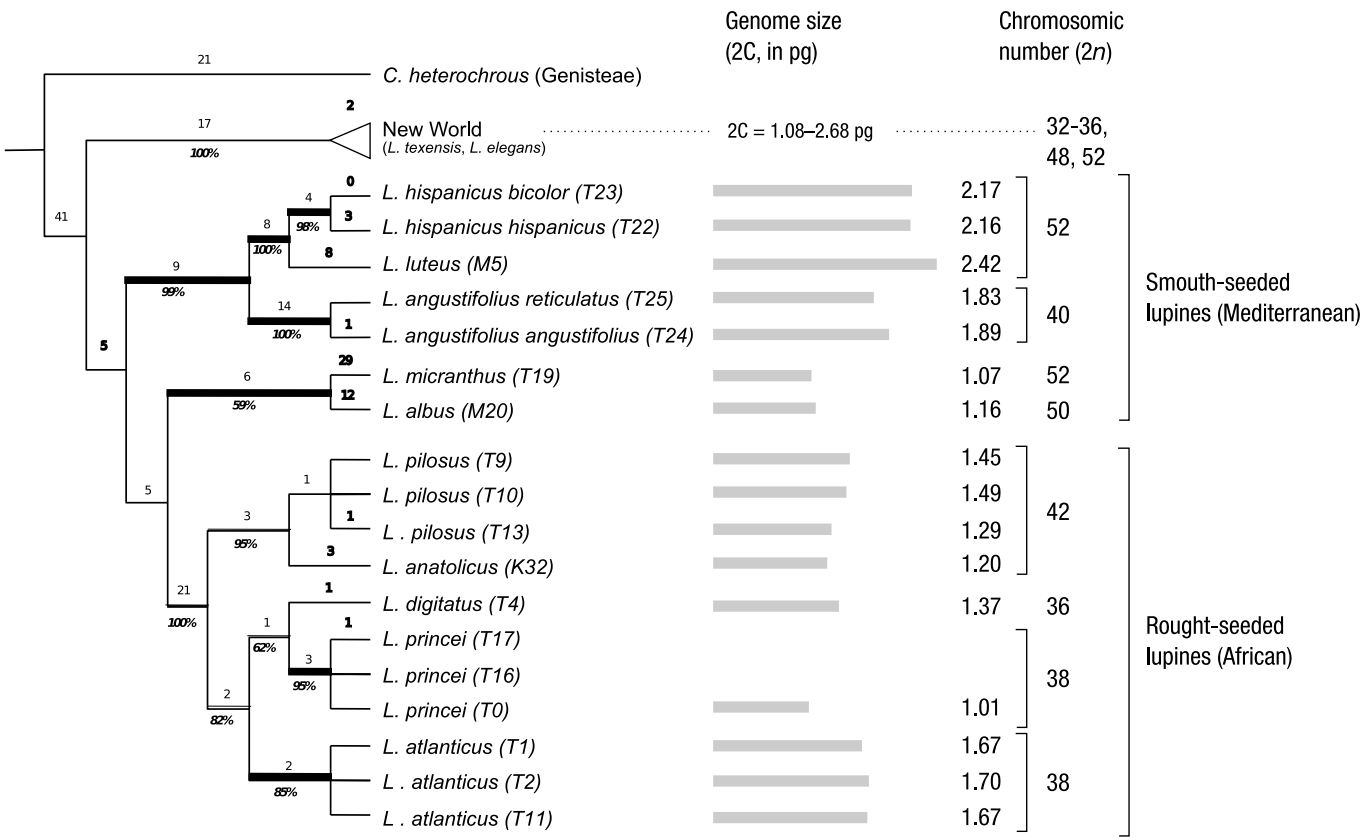

FIG. 7.1 — Phylogeny of *Lupinus* species, genome size variations and number of chromosomes. Parsimony analysis obtained from a combination of ITS-nrRNA and ETS-nrRNA data sets. Some values were taken from other studies: genome size for *L. albus* and New World species (Naganowska *et al.*, 2003), genome size for *L. anatolicus* (Obermayer *et al.*, 1999) and number of chromosomes for New World species (Conterato & Schifino-Wittmann, 2006).

TAB. 7–2 — List of reverse transcriptase sequences from GenBank included in this study to represent retrotransposons from other Fabaceae and Solanaceae.

Species	**Ty1-*copia***	**Species**	**Ty3-*gypsy***
Cicer arietinum	AJ411775	*Glycine max*	AC186736
Glycine max	D12839	*Gossypium herbaceum*	EU098784
Nicotiana tabacum	D83003	*Lotus japonicus*	AP004896
Petunia hybrida	M94487	*Medicago truncatula*	AC144805; CT033763
Solanum lycopersicum	AC171731	*Pisum sativum*	AF378049
Solanum tuberosum	AJ228808	*Populus trichocarpa*	AC216590
Vicia sativa	AJ239509	*Solanum lycopersicum*	AC212439
Vigna radiata	AY684683	*Vigna radiata*	AY683021; AY683026
Vigna unguiculata	Y12763		

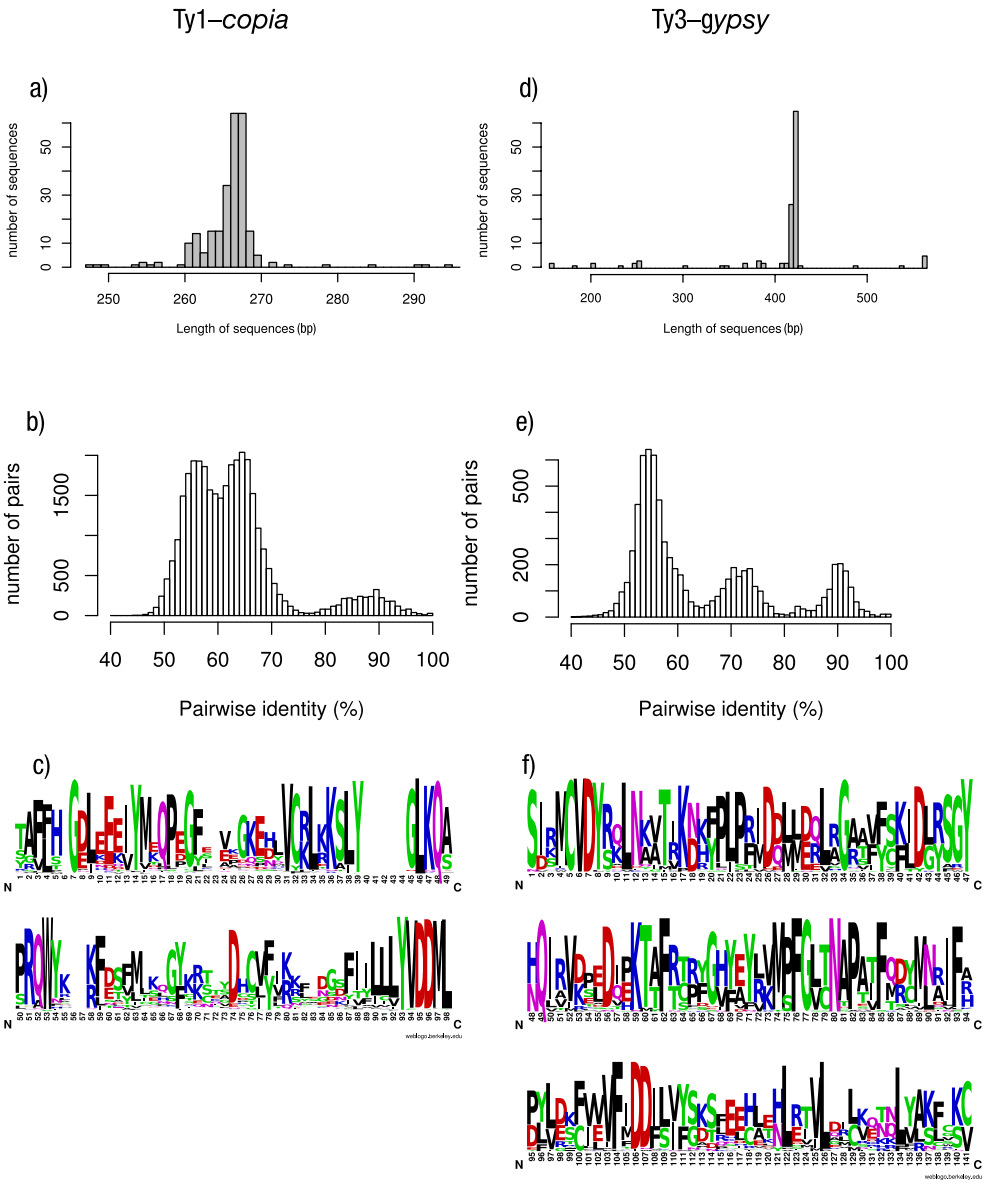

FIG. 7.2 — Characteristics of Ty1-*copia* and Ty3-*gypsy rt* sequences amplified in lupines: length distribution (a,d), pairwise-identity distribution (b,e) and amino acid frequencies of putatively functional sequences (c,f).

			Ty1-copia		Ty3-gypsy		
Origin	**species**	**code**	**clones/pop**	**Total/taxa**	**clones/pop**	**Total/taxa**	**Total**
New World	*L. affinis*	N20	7	7			7
	L. bracteolaris	S80	16	16			16
	L. concinnus	N19	8	8			8
	L. diffusus	K35			5	5	5
	L. elegans	S33	4	4			4
	L. hirsutissimus	N85	1	1			1
	L. mutabilis	MU23	10	10			10
	L. nanus	N42	9	9			9
	L. paraguarensis	BZ3	6	6			6
	L. polyphillus	T26	1	1			1
	L. texensis	N45	9	9			9
Old World	*L. albus albus*	M20	2	2	14	14	16
	L. anatolicus	K32	6	6			6
	L. angustifolius angustifolius	T24	2	2	3	3	5
	L. angustifolius reticulatus	T25	10	10			10
	L. atlanticus	T1	7	22	11	25	18
		T2	5		7		12
		T11	10		7		17
	L. consentinii	T15	5	5			5
	L. digitatus	T4	11	11			11
	L. hispanicus hispanicus	T22	7	7			7
	L. hispanicus bicolor	T23	4	4	5	5	9
	L. luteus	M5	1	7	5	16	23
		T20	4		11		
		T21	2				
	L. mariae-josephi	MJ 1	14	14			14
	L. micranthus	T19	11	13			13
		T28	2				
	L. palaestinus	T14	7	7			7
	L. pilosus tassilicus	A641	13	13			13
	L. pilosus	T6					
		T9	10	10			10
		T13			7	7	7
	L. princei	T16	6	11	13	28	39
		T17	1		4		
		T0	4		11		
Genistoids	*Anarthrophyllum cumingii*	201	3	3			3
	Argyrolobium uniflorum	G25	7	7			7
	Chamaecytisus mollis	C84	8	8			8
	Crotalaria podocarpa	K50	7	7	11	11	18
	Cytisus heterochrous	G8			6	6	6
	Genista tinctoria	G56	5	5			5
	Thermopsis rhombifolia	G46	6	6			6
	Ulex parviflorus	G24	9	9			9
				260		**120**	**380**

FIG. 7.3 — Number of clones obtained for each of the 38 accessions sampled (*Lupinus* and outgroups).

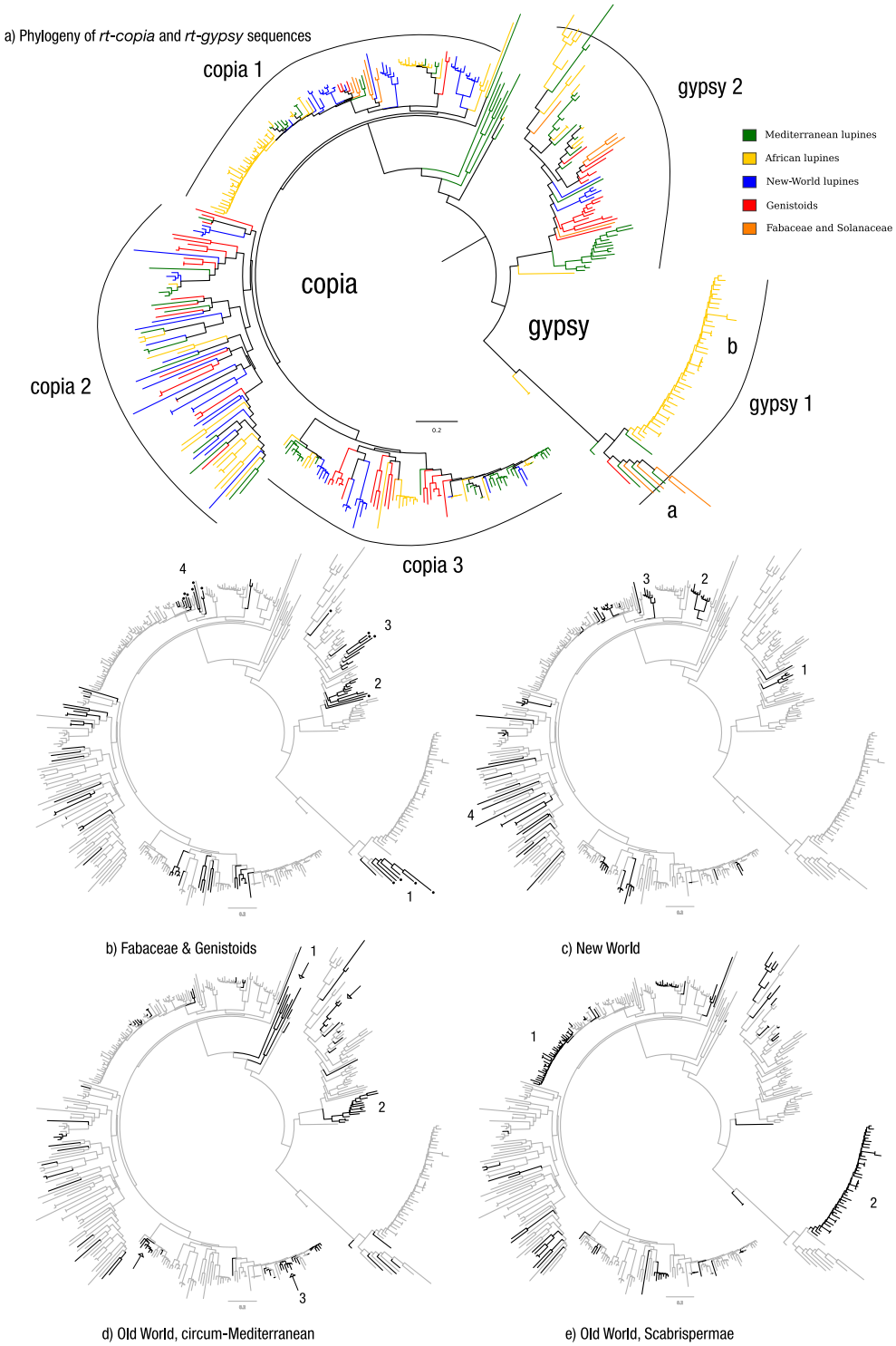

FIG. 7.4 — Neighbor Joining tree of 260 *copia* and 120 *gypsy* reverse transcriptase (*rt*) sequences obtained from 27 species of lupines and 7 Genistoids through amplification, cloning and sequencing. Additional RT sequences from GenBank, representing other Fabaceae and Eudicots (indicated by black dots in the part b of the figure), were included in this study for comparison.

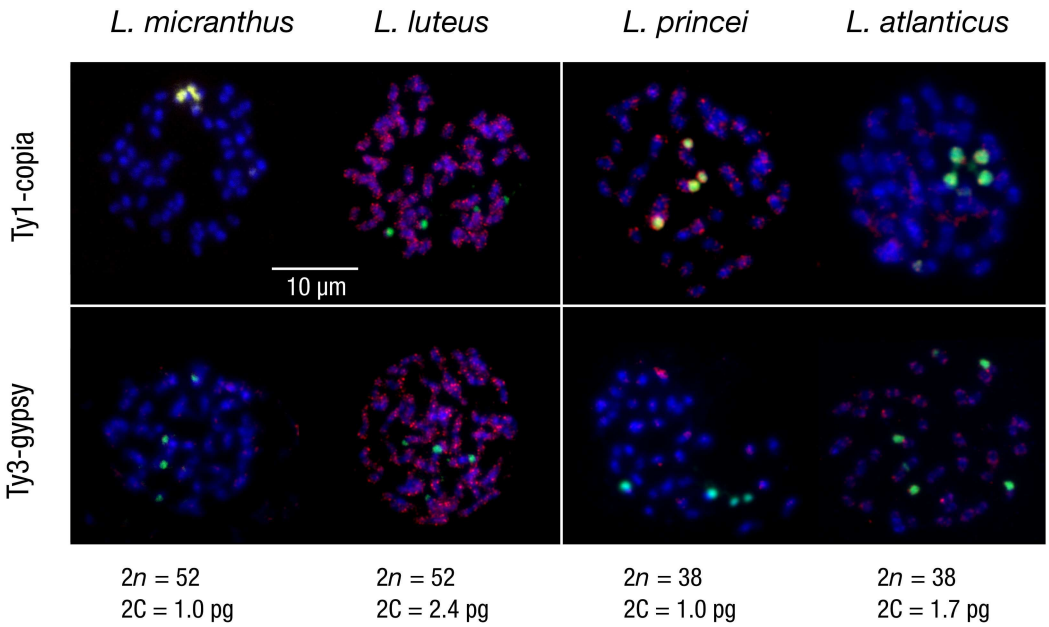

FIG. 7.5 — Fluorescence in situ hybridization of Ty1-*copia* and Ty3-*gypsy* reverse transcriptase probes on two couples of Old World lupine species with remarkable genome size differences. Chromosomes are in blue, 45S rRNA loci are in green, *rt* probes are in red. The scale is the same for all images.

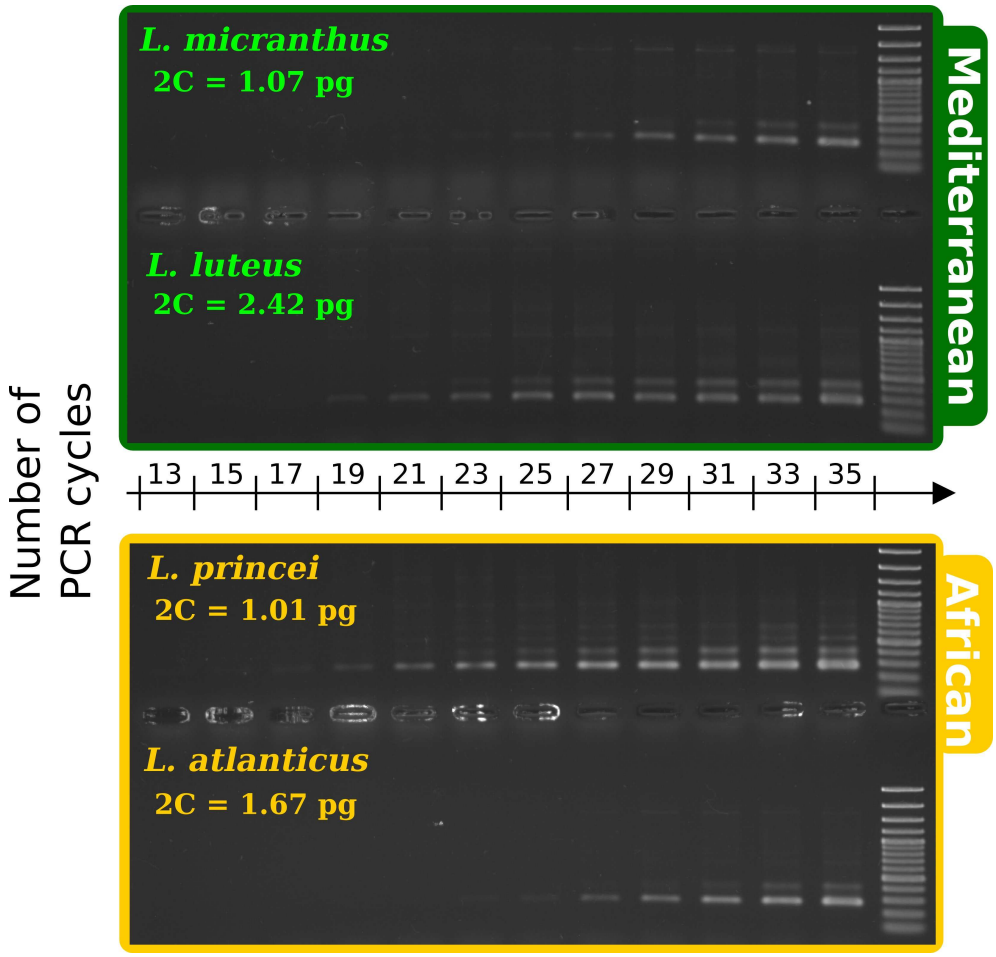

FIG. 7.6 — Semi-quantitative PCR-based estimates of the number of *copia* elements in four Old World lupines. The more important is the difference in the kinetic of amplification between the two species, the more different is the copy-number ratio.

8

Analyse génomique de la région *SymRK*

DU FAIT de leur capacité unique à s'associer à des bactéries symbiotiques fixatrices d'azote, les légumineuses représentent une source d'azote biologique considérable pour le fonctionnement des écosystèmes naturels et agricoles (protéines, métabolites secondaires). Cette famille de plantes fait ainsi l'objet d'une attention particulière de la communauté scientifique internationale pour l'exploration de la structure, des fonctions et de l'évolution de leurs génomes. À ce jour, les travaux restent principalement centrés sur des plantes modèles ou d'intérêt agronomique qui se rapportent principalement à des clades parmi les plus dérivés de la sous-famille des papilionacées (Galegoideae ; Millettioideae). Nous nous proposons d'étendre cette exploration à des organismes non-modèles, ayant des propriétés biologiques et écologiques particulières, et représentant des lignées de légumineuses d'origine plus anciennes, telles que les génistoidées.

Dans ce cadre, nous présenterons une analyse comparative de la région génomique du gène *SymRK*, qui joue un rôle clé dans les processus d'association symbiotique plantes/micro-organismes (mycorhizes et nodules), chez des représentants modèles et non modèles de différentes lignées de légumineuses papilionacées : gallegoïdes (*Medicago truncatula* et *Lotus japonicus*), millettioïdes (*Glycine max*) et génistoïdes (*Lupinus angustifolius*). Nos résultats permettent une première évaluation du degré de conservation de la synténie et de la micro-colinéarité à l'échelle d'une région localisée du génome chez des lignées phylogénétiques d'âges différents.

Ne disposant pas de ressources génomiques pour le lupin, nous avons cherché à évaluer l'impact des éléments transposables dans une région particulière du génome, en mettant à profit l'existence d'une banque BAC disponible dans au laboratoire de Bogdan Wolko, à Poznán en Pologne (Kasprzak *et al.*, 2006). Pour cela, nous avons choisi de cibler et séquencer une région contenant un gène d'intérêt, le *SymRK*, ayant fait l'objet d'analyses phylogénétiques et moléculaires au chapitre 6. Nos objectifs sont les suivants :

1. évaluer l'impact des éléments transposables à une échelle locale du génome de *Lupinus angustifolius* à partir du séquençage et de l'annotation d'un BAC ;

2. comparer cette région annotée aux séquences homologues disponibles pour les légumineuses-modèles.

Ce travail fera prochainement l'objet d'un article intitulé *Conserved synteny and divergence in the SymRK genomic region among different Papilionoid lineages* (Frédéric Mahé, Bogdan Wolko, Marie-Thérèse Misset & Abdelkader Aïnouche). Nous présentons ci-dessous les résultats préliminaires, la méthodologie ayant été décrite page 87.

8.1 Annotation du BAC

L'assemblage du BAC a nécessité 1 536 séquences de 500 à 800 paires de bases aboutissant à une séquence finale de 116 636 nucléotides (couverture moyenne de $12,7\times$). La répartition en base est la suivante : *A* (39 512), *C* (19 661), *G* (19 286) et *T* (38 177). La teneur en $G + C$ est donc de 33,4 %, ce qui est comparable aux valeurs rencontrées chez *Arabidopsis thaliana* (35 %) ou *Vitis vinifera* (36,2 %).

Les logiciels utilisés pour detecter les régions exoniques — GENSCAN et FGENESH — ont prédit 29 gènes et 127 exons, se répartissant de la façon suivante entre les deux brins : 14 gènes et 45 exons sur le brin direct, 15 gènes et 82 exons sur le brin complémentaire. La taille moyenne est de 159 pb pour les exons, et de 144 pb pour les introns. La confrontation de ces gènes prédits avec les bases de données publiques a montré que nos séquences trouvent leurs homologues principalement chez *Vitis vinifera*, *Oryza sativa* L. ssp. *japonica*, *Arabidopsis thaliana* et dans une moindre mesure chez *Medicago truncatula*.

La séquence du BAC a également été soumise à deux logiciels dédiés à l'identification d'éléments répétés : REPEATMASKER et CENSOR. Le logiciel REPEATMASKER détecte 11 portions d'éléments à LTR (8 908 pb au total, soit 7,64 % du BAC) ainsi que 16 petites répétitions et 54 régions de faibles complexité (homopolymères, 2 439 pb au total, soit 2,09 % du BAC). Le logiciel CENSOR détecte 42 portions d'éléments à LTR (14 071 pb, soit 12,06 % du BAC) ainsi que 19 éléments de classe II (2 147 pb, soit 1,84 % du BAC). La combinaison de ces résultats, l'utilisation d'autres bases de données spécialisées comme *Plant Repeat Databases*[1] (Ouyang & Buell, 2004) et l'utilisation de techniques comme l'annotation transitive permettent de compléter cette étape d'identification. Les rétroéléments Ty1/*copia* et Ty3/*gypsy* constituent au final l'essentiel des éléments identifiables, tant par leur nombre que par la longueur des séquences, et couvrent près de 25 % de la séquence du BAC (voir Fig. 8.1 page 156). Ces éléments ne sont pas répartis uniformément le long de la séquence, mais forment des régions riches en éléments transposables, séparant des régions riches en gènes.

La première région riche en gènes s'étend sur environ 45 kb. Elle contient 6 gènes identifiés et deux gènes inconnus. Elle contient également quelques séquences répétées de petite taille, annotées comme des transposons. Le premier gène est une β-D-xylosidase similaire à celles rencontrées chez les autres légumineuses. Le deuxième gène est le gène *SymRK*. Étant au cœur de notre travail, le *SymRK* est utilisé ici comme

1. `http://plantrepeats.plantbiology.msu.edu/search.html`

référent pour orienter le BAC. Le brin sur lequel se trouve le *SymRK* est considéré comme le brin normal. Les orientations relatives des autres gènes sont indiquées sur la figure 8.1 par des rectangles blancs (sens) ou noirs (antisens). Le troisième gène, *Brick1*, est annoté comme un composant du *scar regulatory complex*. Le quatrième gène code pour l'enzyme oligopetidase A, impliquée dans la maturation des protéines, et le cinquième gène code pour l'enzyme glycine-hydroxyméthyltransférase.

La seconde région (d'environ 22 kb) est plus réduite que la première et ne contient que trois gènes, dont deux sont identifiés. Le premier gène est annoté comme une *dof zinc finger protein*, un facteur de transcription impliqué dans la germination. Le deuxième gène, composé d'un grand nombre d'exons (3,657 bp, 23 exons), code pour une *kinesin motor*.

Les deux régions riches en gènes sont séparées par une région d'environ 40 kb riche en éléments répétés, comme indiqué par les nombreux rubans joignant les paires de séquences similaires. À noter que si les fragments de rétroéléments composent l'essentiel de cette région, aucun élément à LTR complet n'est présent. La deuxième région riche en éléments transposables couvre les 10 000 dernières paires de bases de la partie 3' du BAC, et comporte 14 exons non-identifiés mais présentant des similitudes avec des rétrotransposons de la première région riche en éléments transposables. En dehors de ces régions riches en éléments transposables, 21 fragments sont présents dans les espaces intergéniques ou dans les introns de certains gènes, comme par exemple l'intron 9 du *SymRK*.

8.2 Comparaison avec des régions génomiques homologues

Afin de replacer la région *SymRK* obtenue pour *Lupinus angustifolius* dans une perspective évolutive, des régions homologues ont été recherchées dans les génomes de fabacées en cours de séquençage (*Lotus japonicus*, *Medicago truncatula* et *Glycine max*). La figure 8.2 page 157 montre la comparaison de la région du gène *SymRK* présente chez *Lupinus angustifolius* avec la région homologue présente chez *Medicago truncatula* et *Glycine max*. Les deux régions riches en gènes observées chez *Lupinus* sont également présentes chez *Medicago* et *Glycine*, avec une bonne conservation de la synténie. Par contre, l'intervalle entre ces deux régions n'est pas occupé par des éléments transposables chez ces deux taxons. Chez *Glycine* et *Medicago*, il s'agit d'une région riche en gènes, et notamment en lectines, codant des protéines de reconnaissances d'oses ou d'oligosides et abondantes dans les graines de légumineuses.

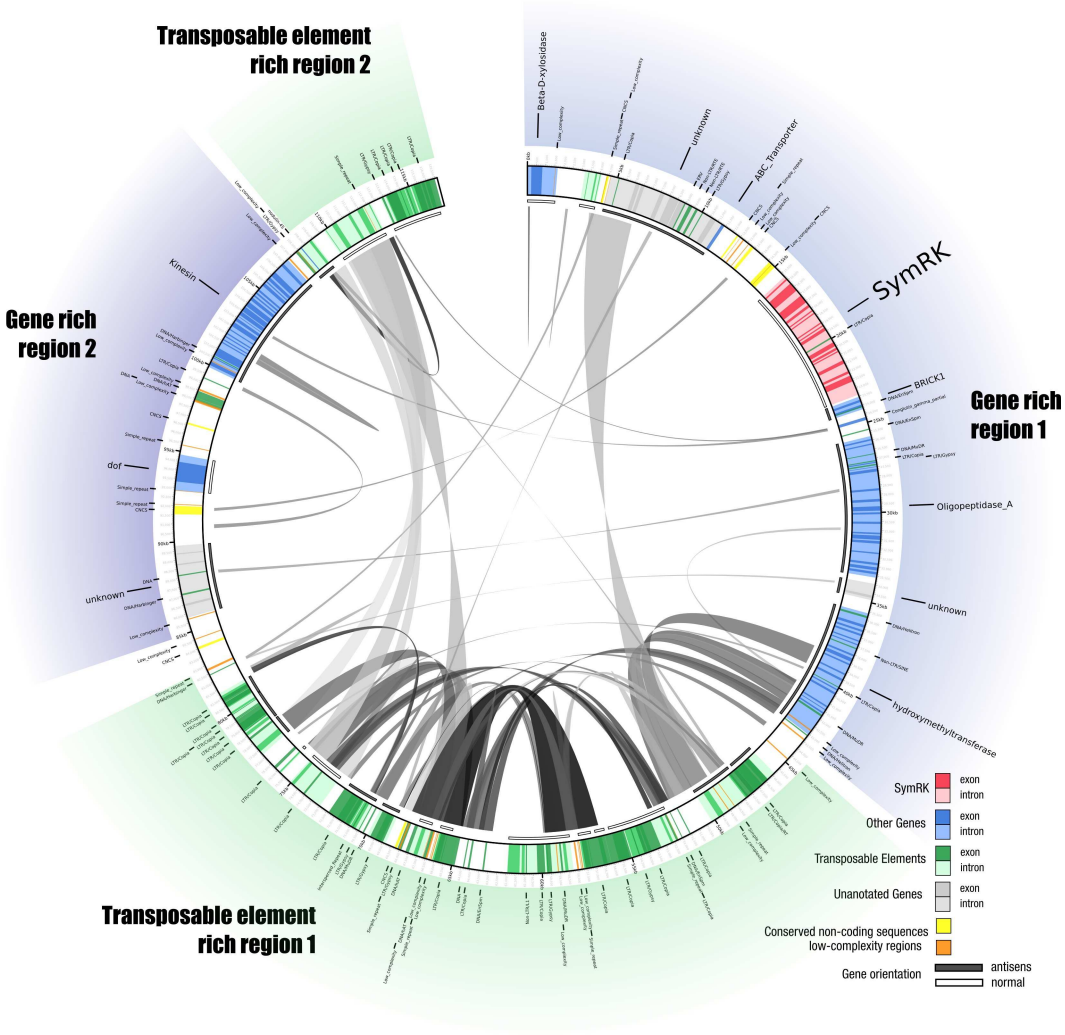

FIG. 8.1 — Annotation de la première séquence génomique de *Lupinus angustifolius*. Cette séquence de 116 kb compte huit gènes identifiés (en bleu) dont le gène d'intérêt *SymRK* (en rouge). Les séquences répétées au sein du BAC sont indiquées par des rubans gris, l'intensité du gris reflétant le degré de similarité entre les régions répétées. Les gènes non-identifiés sont en gris et les éléments transposables sont en vert. Les gènes identifiés sont groupés dans deux îlots (*gene rich regions* 1 et *gene rich regions* 2), séparés par des régions riches en éléments transposables (*transposable elements rich regions* 1 et *transposable elements rich regions* 2).

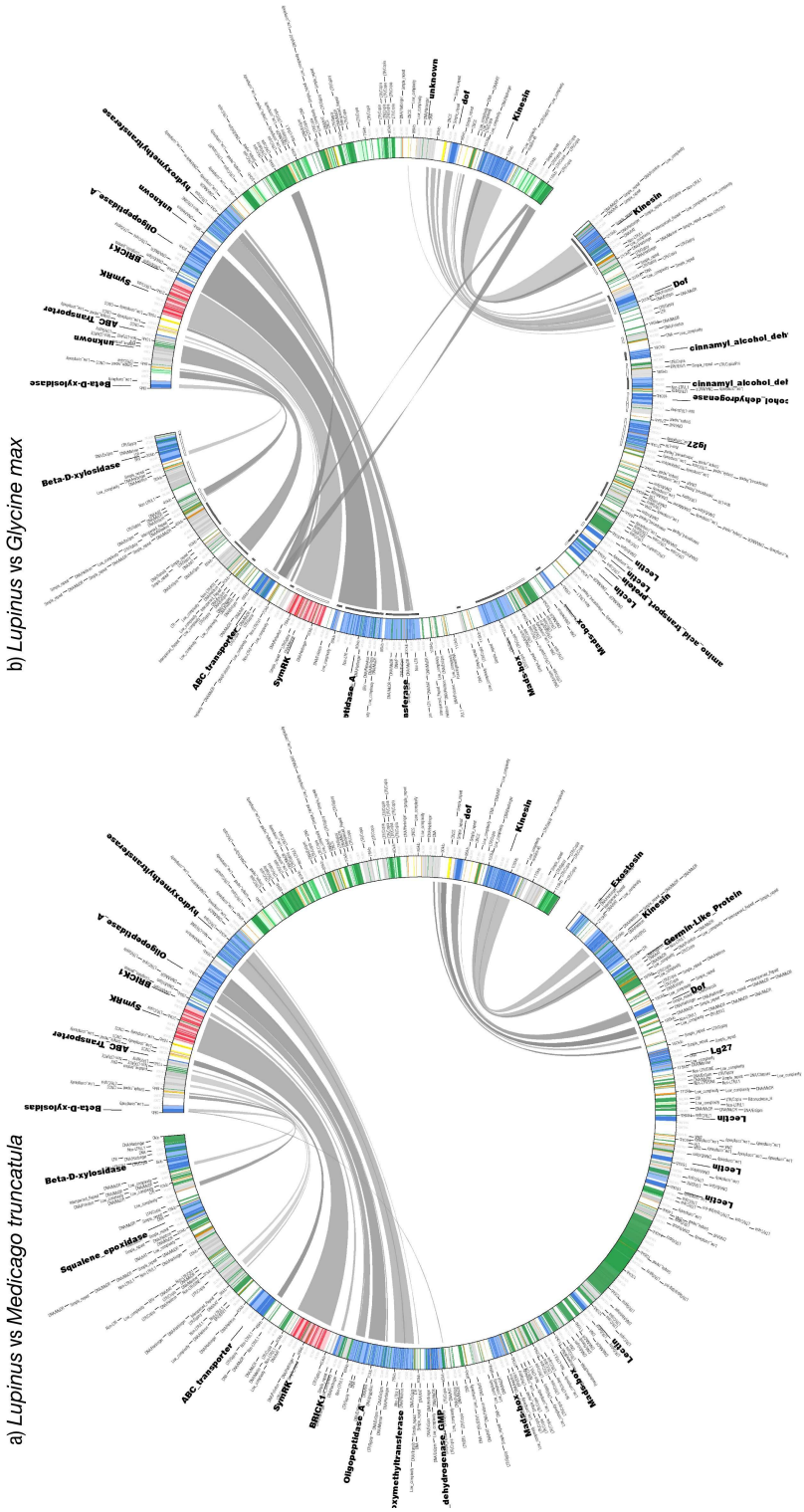

FIG. 8.2 — Comparaison de la région du gène *SymRK* présente chez *Lupinus angustifolius* avec la région homologue présente chez a) *Medicago truncatula* et b) *Glycine max*. La séquence de *Lupinus* est placée sur la partie droite de chaque figure et les rubans gris relient entre-elles les régions conservées chez les deux taxons. On distingue nettement deux régions conservées.

La recherche de régions homologues a ensuite été étendue au génome de *Populus trichocarpa* (voir Fig. 8.3 et Fig. 8.4 page ci-contre). Malgré la distance évolutive séparant les différentes lignées de papilionoïdées, la structure de la région comportant le gène *SymRK* est très conservée. Cette synténie est également retrouvée à l'extérieur des papilionoïdées chez *Populus*, ce qui suggère l'existence d'une contrainte sélective forte s'exerçant sur cette région. Chez *Glycine*, *Medicago* et *Lotus*, la région décrite comme riche en éléments transposables chez *Lupinus* est une région riche en gène. Cet important évènement de restructuration de la région flaquant le « groupe *SymRK* » pourrait être due à l'activité d'éléments transposables. Toutefois, en l'absence de données disponibles pour un groupe externe au papilionoïdées, cette différence remarquable ne peut pour l'instant pas être polarisée (s'agit-il d'un caractère ancestral ou d'un caractère dérivé ?). Ce résultat souligne le besoin d'élargir le champ des études génomiques vers des espèces non-modèles représentant plusieurs lignées des papilionoïdées.

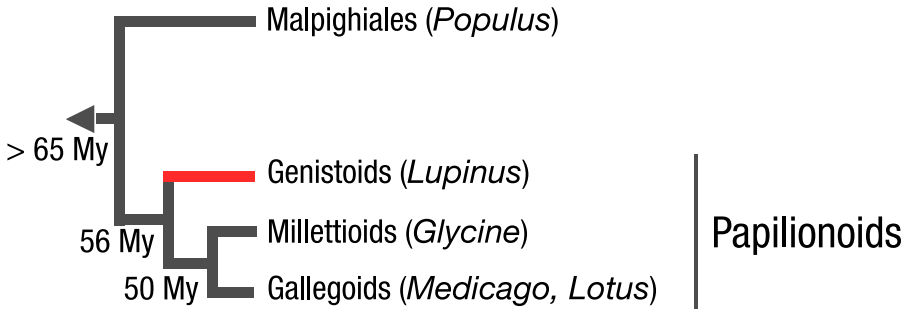

FIG. 8.3 — Position du genre *Lupinus* par rapport à quatre espèces modèles en cours de séquençage.

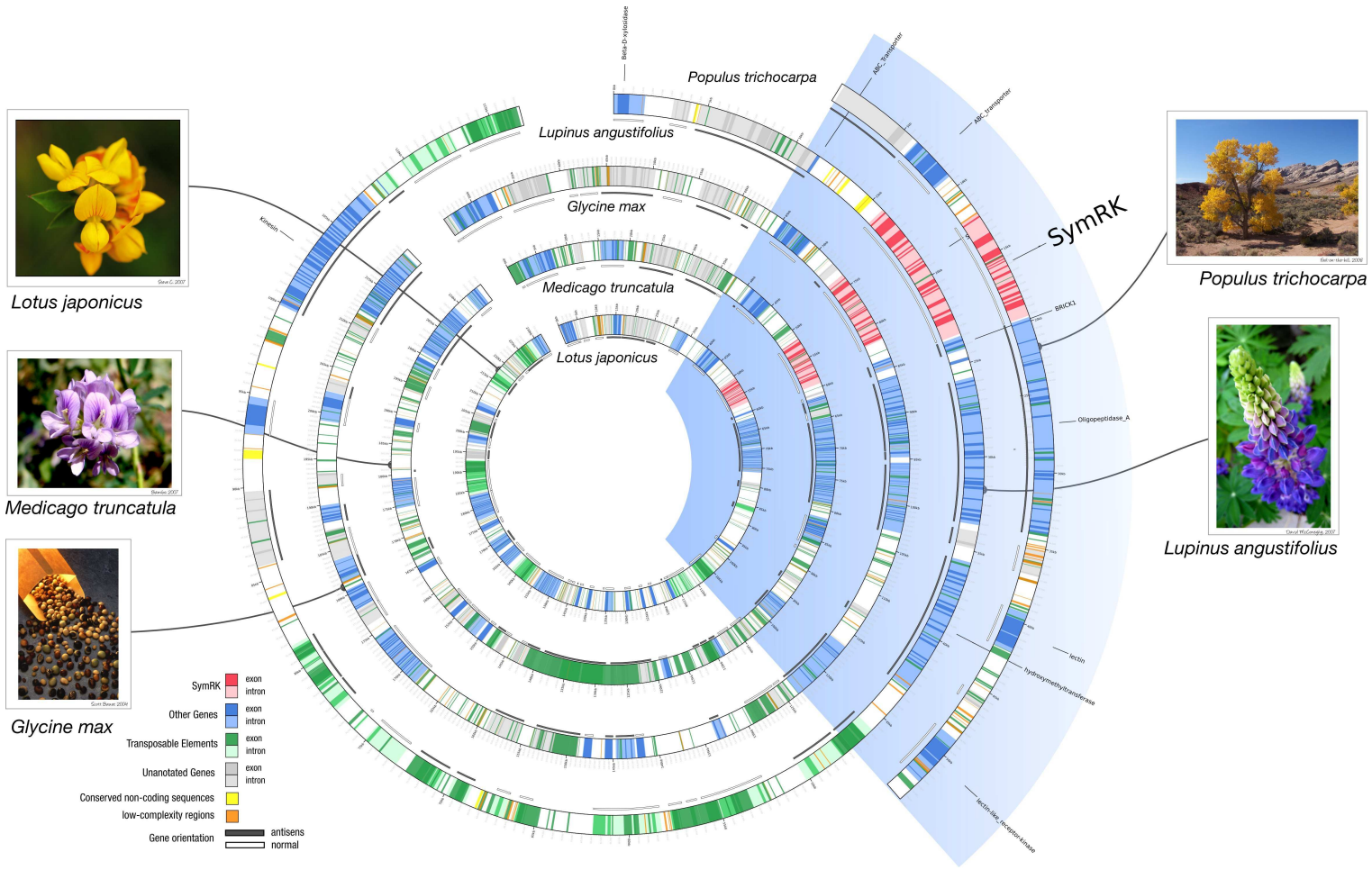

FIG. 8.4 — Comparaison de la région *SymRK* de *Lupinus angustifolius* avec trois régions homologues issues d'autres papilionoïdées (*Lotus*, *Medicago* et *Glycine*) et avec une région plus courte issue de *Populus trichocarpa*. Le gène *SymRK* est indiqué en rouge et la région conservée dont il fait partie est soulignée par un fond bleu.

Quatrième partie

Discussion générale et conclusion

9

Bilan et perspectives

« Voici le dernier travail et la fin de longs voyages. »
Virgile (70–19 avant J.-C.) *L'Énéide*, livre III[1].

AU COURS DE CE TRAVAIL, nous avons examiné le rôle des éléments transposables dans la variation de la taille des génomes chez les lupins, en relation avec leur diversification et leur distribution dans des conditions écogéographiques variées. Nous avons mené pour cela plusieurs approches en parallèle, visant à (1) améliorer notre compréhension de l'histoire évolutive des espèces actuelles du genre *Lupinus* (phylogénies moléculaires), (2) identifier les catégories d'éléments transposables potentiellement impliquées dans les variations de taille des génomes entre les différentes lignées de lupins, et (3) séquencer une première région génomique centrée sur le gène d'intérêt *SymRK* — gène impliqué dans les interactions symbiotiques avec les mycorhizes et les bactéries fixatrices d'azote — et analyser à l'échelle de cette région l'impact structural et la distribution des éléments transposables.

Ces approches ont fait intervenir différentes techniques moléculaires (clonage, séquençage, hybridation *in situ*), plusieurs méthodes de phylogénie et d'évolution moléculaire, et d'analyses bioinformatiques (annotation, génomique comparative). Les deux nouveaux jeux de données provenant de gènes nucléaires répétés (famille des gènes ribosomiques) et de gènes en simple copie (*SymRK*) que nous avons obtenu chez les lupins ont permis de clarifier les relations phylogénétiques entre plusieurs lignées dont les positions étaient mal résolues jusque là : monophylie des lupins du Nouveau Monde, liens particuliers entre la lignée des lupins unifoliolés de Floride et les lupins de l'Ancien Monde, et diversification des lignées circum-méditerranéennes et africaines de l'Ancien Monde.

Un résultat original concerne la position phylogénétique d'une espèce nouvellement décrite dans la région méditerannéenne (*L. mariae-josephi*), qui montre une divergence ancienne des autres espèces de l'Ancien Monde, et des liens particuliers avec les lupins unifoliolés du Nouveau Monde. Les topologies conflictuelles obtenues à par-

1. *Hic labor extremus, longarum haec meta viarum.*

tir de différents gènes suggèrent l'éventualité d'événements passés d'hybridation et introgression ayant affecté certaines lignées. Les relations phylogénétiques à la base du genre restent toutefois difficiles à établir, et résultent probablement d'une radiation qui semble confirmée par les analyses phylogénétiques publiées à ce jour. Les perspectives de travail concernant la résolution des nœuds anciens de la phylogénie des lupins devraient élargir l'échantillonnage des gènes nucléaires ; en effet, à ce jour, cinq gènes nucléaires seulement ont été utilisés chez les lupins (dont deux au cours de cette étude) : gènes ribosomiques (régions ITS et ETS), gènes de la famille *cycloideae* (LEGCYC1A, LEGCYC1B), le gène glycérol 3-phosphate acyltransférase (GPAT1), et le gène *SymRK*. Les progrès effectués dans la connaissance des génomes des génistoidées et des lupins devraient permettre le développement de multiples phylogénies d'orthologues.

Les données du gène *SymRK*, qui s'est révélé particulièrement informatif au plan phylogénétique, nous ont également permis d'estimer les dates de formation des différents clades de lupins, situant le dernier ancêtre commun des lupins actuels entre 6 à 11 millions d'années, la divergence entre lupins du Nouveau Monde et de l'Ancien Monde remontant à 4-7 millions d'années. Chez les lupins de l'Ancien Monde, étudiés plus en détail dans cette thèse, nous montrons notamment la diversification ancienne (fin du Miocène, début Pliocène) des lupins méditerranéens et africains, dans lesquels les lupins à « graines rugueuses » sont dérivés des lupins à « graines lisses » de la région méditerranéenne. Un résultat remarquable correspond à la divergence récente (1,7 millions d'années) entre les lupins de l'Est de la Méditerranée (Moyen-Orient et Anatolie) et les lupins africains colonisant les régions arides et désertiques d'Afrique du Nord, à partir desquelles s'est différenciée il y a moins de 0,2 millions d'années *L. princei*, une espèce endémique du Kenya, seule espèce de lupin de l'Ancien Monde adaptée aux conditions équatoriales.

Ces résultats nous permettent d'examiner l'évolution de la taille des génomes dans un contexte phylogénétique relativement bien résolu. En nous focalisant sur les lignées de l'Ancien Monde, nous avons pu montrer une variation importante de la quantité globale d'ADN chez les espèces méditerranéennes à « graines lisses », allant du simple au double à même niveau de ploïdie. L'examen de cette variation à la lumière des liens phylogénétiques indique différents modes d'évolution de la taille des génomes chez le lupin. En effet, dans certains cas, la variation de la quantité d'ADN semble corrélée à la phylogénie, comme chez les lignées sœurs *L. hispanicus* et *L. luteus* ($2n = 52$) qui présentent des tailles de génomes relativement proches (respectivement 2,42 pg et 2,17 pg) et nettement plus importantes que les tailles de génome de *L. micranthus* ($2n = 52$; $2C = 1,07$ pg) et de *L. albus* ($2n = 50$; $2C = 1,16$ pg), phylogénétiquement plus éloignées. Il est intéressant de noter que cette variation s'est accompagnée d'une adaptation à des conditions édaphiques contrastées : large tolérance pour le premier couple d'espèce et restriction aux sols sableux pour le second couple.

Dans d'autres cas, nous avons au contraire noté une variation nette de la taille des génomes entre espèces phylogénétiquement proches, comme l'endémique du Kenya *L. princei* ($2n = 38$, $2C = 1$ pg) et l'endémique du Maroc *L. atlanticus* ($2n = 38$,

$2C = 1,7$ pg), ces deux espèces appartenant au même sous-clade que l'endémique égyptienne *L. digitatus* ($2n = 36$; $2C = 1,4$ pg). Il s'avère donc que sur une courte période de temps évolutif (moins d'un million d'années), ces espèces aient subi d'importants changements génomiques depuis leur différenciation à partir d'un ancêtre commun. La position de *L. princei* dans le groupe indique que la petite taille du génome de cette espèce résulterait d'une perte rapide d'ADN ayant accompagné la colonisation de nouveaux milieux en région tropicale, à partir d'un ancêtre nord-africain. De la même manière, la plus grosse taille du génome de *L. atlanticus* résulterait d'une accumulation d'ADN au cours de sa propre histoire évolutive.

Ces conclusions illustrent une propriété particulièrement importante des modalités d'évolution des génomes des plantes, et plus généralement des eucaryotes, montrées par les approches de phylogénie comparative, révélant la bidirectionnalité de l'expansion/contraction des tailles de génomes au cours de l'histoire des lignées.

Les éléments mobiles de classe I (ou rétroéléments), transposant selon un mode « copier-coller » représentent généralement une fraction importante des séquences impliquées dans les variations de la taille des génomes. Nous avons analysé les séquences de *reverse transcriptase* (*rt*) de deux familles de rétroéléments Ty1/*copia* et Ty3/*gypsy* chez les lupins et retracé leur dynamique évolutive. Nos résultats ont montré une grande diversité des séquences *rt* chez les lupins. Toutefois si certaines sous-familles s'avèrent ubiquistes (et donc plus anciennes) chez les génistoidées auxquelles les lupins se rattachent, certains groupes de rétrotransposons paraissent amplifiés spécifiquement chez certaines lignées de lupins, soit chez les espèces méditerranéennes soit chez les espèces africaines. Nos estimations relatives des nombres de copies de ces familles d'éléments (par PCR semi-quantitative et hybridation *in-situ*) montrent que les éléments *copia* et *gypsy* contribuent de façon plus significative aux différences de taille de génome chez les lupins méditerranéens que chez les lupins africains. Ceci indique également l'existence de différents modes et mécanismes d'évolution de la taille des génomes au sein d'un même genre. Les différences de tailles de génomes chez les lupins africains devraient donc être recherchées à l'avenir dans les autres compartiments répétés du génome (éléments de classe II, ADN satellite, ...).

La compréhension des mécanismes impliqués dans l'évolution des génomes de lupins passe par le développement de ressources génomiques, inexistantes à ce jour chez les génistoidées. Au moment où nous avons entamé ce travail, les ressources génomiques dans la famille des papilionoidées étaient extrêmement réduites, les assemblages de séquences génomiques n'étant disponibles qu'à partir de 2008 chez *Lotus*, et plus récemment (2009) chez *Medicago* puis *Glycine*. L'opportunité qui nous a été offerte de collaborer avec une équipe polonaise ayant réalisé une banque BAC pour l'espèce méditerranéenne *Lupinus angustifolius* nous donnait une excellente occasion d'entamer un programme de séquençage sur un génome de lupin, par ailleurs inscrit sur la liste des représentants prioritaires de lignées de légumineuses à analyser dans les projets de séquençage massif.

Nous avons séquencé et annoté 116 kb de la région contenant le gène *SymRK*, qui a été comparée aux régions orthologues des plantes modèles relativement proches dis-

ponibles dans les bases de données (*Lotus*, *Medicago*, *Glycine*, *Populus*). Nous avons ainsi pu détecter une restructuration majeure des domaines flanquant la région riche en gènes contenant *SymRK*, après la divergence des génistoidées des autres lignées plus dérivées des papilionoidées (estimée à 50-60 millions d'années). Chez *Lupinus angustifolius*, les éléments transposables représentent environ 25 % de la région analysée, dont essentiellement des fragments de rétrotransposons à LTR *(Long Terminal Repeats)* de type *copia* et *gypsy*, confirmant ainsi la forte implication de ces familles d'éléments détectée par nos précédentes analyses.

Les possibilités nouvelles offertes par les technologies de séquençage de deuxième et troisième génération devraient permettre d'augmenter rapidement la quantité de données chez différentes espèces de lupins à « gros génome » et « petit génome », qui re-situées dans leur contexte phylogénétique devraient amener de nouveaux éclairages sur les mécanismes génomiques ayant accompagné la diversification des lupins.

En conclusion, ce travail nous a permis d'appréhender l'évolution d'un groupe biologique à différents niveaux, depuis l'histoire des organismes constituant les principales lignées de ce genre extrêmement diversifié (plusieurs centaines d'espèces distribuées sur différents continents, et occupant une grande diversité de milieux), à la dynamique individuelle des séquences (histoire du gène *SymRK*, histoire de familles d'éléments transposables) et des génomes (dynamique structurale, évolution de la synténie). Ces données, qui permettent une meilleure compréhension des mécanismes accompagnant la spéciation et l'adaptation, pourront également trouver leur prolongement appliqué au regard des multiples utilisations des lupins. En région méditerranéenne, *L. angustifolius*, *L. luteus* et *L. albus* sont en effet cultivées. Leur impact écologique en font des espèces d'intérêt particulier pour les sols à enrichir ou les milieux primaires ouverts à la colonisation ; un atout particulier de la nouvelle espèce *L. mariae-josephi* étant sa capacité à coloniser les sols calcaires pauvres. Enfin, certaines espèces comme *L. angustifolius* montrent de fortes capacités de dépollution (dégradation de l'atrazine et accumulation de métaux lourds) et sont utilisées en phytoremédiation.

Annexes

A

L'énigmatique *Lupinus mariae-josephi*

En 2004, fait rare, une nouvelle espèce de lupin est décrite en Espagne. Les graines récoltées à la fin des années 1980 germent difficilement et la population d'origine est introuvable, détruite par l'extension d'une carrière. Les botanistes de la région entament en 2006 une fouille des reliefs environnants et découvrent trois nouvelles populations. Cette nouvelle espèce présente des caractéristiques intéressantes, comme la capacité à pouvoir pousser sur des sols calcaires (pH basique)[1]. À partir de graines anciennes et d'échantillons des trois populations découvertes en 2006, nous avons réalisé une série d'analyses dans le but de clarifier le statut de cette nouvelle espèce.

Lupinus mariae-josephi possède des caractères morphologiques singuliers la distinguant des autres lupins. Cependant, certains caractères la rapprochent des lupins méditerranéens *L. luteus* et *L. hispanicus*. Comme ces derniers, *L. mariae-josephi* possède $2n = 52$ chromosomes, caractère également partagé par les lupins du Sud-Est de l'Amérique du Nord. La micromorphologie de la surface des graines est habituellement un caractère diagnostic pour les taxons de l'Ancien Monde, mais *L. mariae-josephi* possède une micromorphologie intermédiaire entre celle des lupins africains à graines rugueuses et celle des lupins méditerranéens à graines lisses. Les analyses moléculaires basées sur les espaceurs transcrits des gènes répétés ARNr (ITS et ETS) et sur le gène nucléaire *LEGCYC1A* placent *L. mariae-josephi* avec les lupins de l'Ancien Monde, mais sans mettre à jour de relation claire avec une espèce en particulier. La résolution définitive de la position phylogénétique de *L. mariae-josephi* nécessitera l'utilisation de marqueurs supplémentaires et devrait permettre une meilleure compréhension de l'histoire évolutive du genre *Lupinus*.

Ces résultats seront publiés par la revue *Genetic Resources and Crop Evolution*.

1. Certaines lignées de lupins ont été sélectionnées pour leur capacité à pousser sur des sols très acides (pH $= 4,79$) ou légèrement alcalins (pH $= 7,92$) (Mihailović, Hill, Lazarević *et al.*, 2008 ; Mihailović, Hill, Ćupina & Vasiljević, 2008).

New data and phylogenetic placement of the enigmatic Old World lupine *L. mariae-josephi* H. Pascual

Frédéric Mahé[1], Higinio Pascual[2], Olivier Coriton[3], Virginie Huteau[3], Albert Navarro Perris[4], Marie-Thérèse Misset[1], Abdelkader Aïnouche[1]

[1]UMR CNRS 6553 Ecobio, Université de Rennes-1, Campus scientifique de Baulieu Bât 14A, F-35042 Rennes cedex, France.
[2]Instituto Madrileño de Investigación Agraria y Alimentaria (IMIA), Apartado 127, E-28800 Alcalá de Henares, España.
[3]UMR118 INRA-AgroCampus Rennes INRA Centre de Rennes BP 35327 F-35653 Le Rheu Cedex, France.
[4]Centro de Investigación y Experimentación Forestal (CIEF). Avda. Comarques del País Valencià, 114. E-46930 Quart de Poblet, España.

Lupinus mariae-josephi is an intriguing lupine species recently discovered in the Mediterranean region. New data from seed coat micromorphology, cytology, and DNA sequences were generated in order to extend our knowledge on this species and to examine its evolutionary relationships within *Lupinus*. This species shows morphological similarities with the Mediterranean smooth seeded species of sections *Micranthi* and *Lutei*. It shares the same chromosome number $2n = 52$ with the latter Old World taxa, but also with unifoliolate lupines from Florida. Besides, *L. mariae-josephi* exhibited a seed coat micromorphology intermediate between the rough and the smooth seed coat types. Phylogenetic analyses using ITS and ETS nrDNA spacers, and the *LEGCYC1A* locus supported *L. mariae-josephi* as a distinct Old World lineage, but without clear placement. Unexpectedly, *LEGCYC1A* data indicated phylogenetic affinities between *L. mariae-josephi* and the unifoliolate North American lupine, *L. villosus*, whereas conflicting data from ETS sequences placed *L. villosus* close to the New World lupine clade. This finding suggests that the North American lineage of unifoliolate lupines, which represents a key link between New and Old World lupines, might have experienced reticulate and/or recombination processes. All together, the data highlight the enigmatic status of *L. mariae-josephi* and the complexity of the early evolutionary history of *Lupinus*. Moreover, *L. mariae-josephi* opens new perspectives for ecological and agronomic interests, as it represents the sole lupine species that only grows in poor basic soils, while almost all others occur in acid to neutral soils.

Introduction

The discovery of a new angiosperm species is not a frequent event in Western Europe, one of the best-known areas in the world. In 2004, one of us described *Lupinus mariae-josephi* (Pascual 2004), a new lupine species discovered in a restricted area of eastern Spain, in the region of Valencia. This description was based on seeds collected in the late 1970s and kept and grown in the *Centro de Experimentación Agraria de Carcagente* (Valencia).

Based on a careful morphological comparison with other lupines, this new form was shown to differ from other Mediterranean and American taxa with respect to important diagnostic features, such as the banner position, the flower and the inflorescence structure. Moreover, as no clear relationships could be established with the other lupines growing in Spain (Castroviejo and Pascual 1999), it has been regarded as a new lupine species. As quarrying activity in the 1980s destroyed the original population, the species was thought to be extinct in the wild. In 2006, it was finally re-discovered by botanists from Valencia and one out of the three rediscovered populations of *L. mariae-josephi* is nowadays protected (Navarro Peris et al. 2006).

Lupinus is a wide and diverse genus of Papilionoid legumes with an amphi-Atlantic distribution (Dunn and Gillet 1966; Gladstones 1974, 1998). On one side, hundreds of annual and perennial digitated-leaf lupines with herbaceous and woody habits occur in

the New World. They behave as self- or outcrossing plants and display a generally stable chromosome number of $2n$ = 48 (rarely 96) in most western American lupines, $2n$ = 32 to 36, and 52 in species with digitated or simple leaves growing in southern (Texas) and southeastern (Florida) North America, and southeastern South America (Smith 1944; Phillips 1957; Turner 1957; Dunn and Gillet 1966; Dunn 1984; Planchuelo-Ravelo 1984; Planchuelo and Dunn 1984; Monteiro and Gibbs 1986, Maciel and Schifino-Wittmann 2002; Conterato and Schifino-Wittmann 2006). On the other side, only eleven to twelve species are recognized as native from the Old World (Gladstones 1974, 1998).

Because of their reduced number, their economic potential as protein and nitrogen suppliers (Petterson 1998), and as natural producers of metabolites of great interest for plant defense (Wink 1992; Vega et al. 1996) and human health (Guillon and Champ 2002; Magni et al. 2004; Pilvi et al. 2006), the Mediterranean and African lupines have particularly attracted the attention of the scientific community interested into systematics, genetic resources and applied agricultural engineering. Therefore, they have been fairly well studied from various points of view over the last decades, in order to increase understanding of their biology and ecology, to clarify their taxonomy, their relationships and their evolutionary history (Gladstones 1998; Käss and Wink 1997; Aïnouche and Bayer 1999; Aïnouche et al. 2004; and references therein). The Old World lupines are all annual, herbaceous, predominantly self-crossing, and relatively well characterized on the basis of a variety of criteria from different lines of data: morphology (Gladstones 1974, 1998), micromorphology (Heyn and Herrnstadt 1977; Plitmann and Heyn 1984; Aïnouche and Bayer 2000; Aïnouche et al. 2004), cytogenetics, crossing experiments and flow cytometry (Plitmann and Pazy 1984; Kazimierski 1988; Carstairs et al. 1992; Gupta et al. 1996; Naganowska et al. 2003), isozymes polymorphism (Wolko and Weeden 1990a, 1990b), seed storage protein diversity and serology (Salmanowicz and Przybylska 1994; Przybylska and Zimniak-Przybylska 1995; Aïnouche 1998; Cristofolini 1989), flavonoids (Williams et al. 1983), alkaloids (Nowacki 1963; Wink et al. 1995; Aïnouche et al. 1996; Aïnouche et al. 2004). They represent an aneuploid series with various chromosome numbers ($2n$ = 32, 36, 38, 40, 42, 50 to 52) and variable genome sizes (Naganowska et al. 2003), which is traditionally subdivided into two main groups according to their seed coat microstructure and their geographical distribution (Gladstones 1974, 1998; Heyn and Hernstadt 1977; Plitmann and Pazy 1984). The first group comprises the circum-Mediterranean lupines, which are well differentiated from one another and exhibit a smooth seed coat surface, such as: *L. albus* L., *L. angustifolius* L., *L. luteus* L., *L. hispanicus* Boiss. & Reut., and *L. micranthus* Guss. The second group contains the so-called rough-seeded Mediterranean-African lupines, which are morphologically rather homogeneous and genetically less differentiated, and usually referred to as sect. *Scabrispermae*, including: *L. atlanticus* Gladst., *L. cosentinii* Guss., *L. digitatus* Forssk., *L. palaestinus* Boiss., *L. pilosus* Murray, and *L. princei* Harms. Moreover, successive molecular phylogenetic analyses using various nuclear and plastid DNA sequences have considerably increased our understanding of lupines' taxonomy and systematics in both the Old and the New World (Käss and Wink 1997; Aïnouche and Bayer 1999; Aïnouche et al. 2004; Citerne et al. 2003; Ree et al. 2004; Hughes and Eastwood 2006; Drummond 2008), thus providing an advanced general framework of reference for the genus.

In spite of these significant advances, there is still a need for a careful exploration of some under-investigated areas in the Mediterranean region and North-equatorial Africa in order to complement our knowledge on lupines' diversity in the Old World (Święcicki et al. 1996, Clements et al. 1996). This is well exemplified by the recent discovery of a noteworthy lupine form in southwestern Turkey (*L. anatolicus*; Święcicki et al. 1996), or by the most recent and intriguing case of

L. mariae-josephi, which is the center of interest in this paper. Based upon limited morphological data, it was not possible to clearly situate this species among lupines occurring in the Old World (including the few ones introduced from the New World). As previously shown, several characters and molecular phylogenetic methods have proven to be useful to elucidate the taxonomic status and evolutionary relationships of ambiguous lupines in the Old World (Gladstones 1998; Aïnouche and Bayer 1999, 2000; Aïnouche et al. 2004; and references therein). Using these various sources of data, the objectives of this study were, first, to extend our knowledge of the poorly documented *L. mariae-josephi*, and second, to investigate its placement and affinities within the genus. Here, we present new data on seed coat micromorphology and cytology (chromosome number), and on its phylogenetic placement, as inferred from molecular data.

Material And Methods

Plant material Seeds from the initial population of *L. mariae-josephi*, discovered in the late 1970s at Montserrat de Alcalá (Valencia, Spain), and from the three extant populations recently found in the same region (Navarro Peris et al. 2006) were grown together with representatives of other rough- and smooth-seeded Old World lupines in the green house at the University of Rennes-1 (France), in order to obtain fresh leaf material for molecular analyses.

Seed-coat micomorphology Seed coat surface and hand-cut transverse section of the seed coat were examined using a scanning electron microscope (SEM; JEOL JSM-6301), following the procedure described in Aïnouche and Bayer (2000). At least two seeds per sample were observed and compared to the large database of SEM micromorphological seed-coat patterns, including numerous Old and New World lupine species and populations available from the literature (Heyn and Herrnstadt 1977; Bragg 1983; Plitmann and Pazy 1984; Monteiro 1987; Aïnouche 1998; Aïnouche and Bayer 2000, Aïnouche et al. 2004).

Chromosome counts The chromosome number of *L. mariae-josephi*, was determined from mitotic chromosomes observed on metaphasic cells isolated from root tips of approximately 3 mm. Root tips were pretreated in the dark with 0.04% 8-hydroxiquinoline (2 h at 4°C and 2 h at room temperature), followed by 48 h at 4°C in 3:1 ethanol:acetic acid, and then conserved at –20°C in 70% ethanol until use. After being washed in distilled water for 10 min, then treated 15 min by a 0.01 M pH 4.5 citrate solution, root tips were incubated at 37°C for 30–35 min in a enzymatic mixture (5% cellulase and 1% pectolyase solution). The enzymatic solution was removed and the meristematic root tip was subjected to hypotonic treatment by addition of ultra pure water. After 30 min, root tips are squashed in a drop of 3:1 ethanol:acetic acid. Dried slides were then stained by a drop of 4',6-diamidino-2-phenylindole (DAPI). Cells were viewed with an Olympus BX-51 microscope at 1,000x and digitally documented with a Pixera Penguin 600ES CCD camera.

Molecular phylogenetic analysis Total genomic DNA was extracted from 100 mg of fresh leaf material sampled from a young plant, employing the Nucleospin Plant kit (Macherey-Nagel) and following manufacturer's instructions.

In order to infer the phylogenetic position of *L. mariae-josephi* among its congeners, we amplified and sequenced three nuclear DNA regions in each of the four available samples: two regions from the nuclear ribosomal DNA repeat (Baldwin et al. 1995, Soltis and Soltis 1998), the internal transcribed spacers (ITS region, including ITS1+5.8S+ITS2) and the 3' external transcribed spacers (ETS), on one side; and an orthologous region of the CYCLOIDEA gene loci *LEGCYC1A* (Citerne et al. 2003; Ree et al. 2004; Citerne 2005), on the other side. Both ITS and *LEGCYC1A* regions were previously used for phylogenetic

inference in *Lupinus* (Käss and Wink 1997; Aïnouche et al. 2003; Ree et al. 2004; Hughes and Eastwood 2006), and sequences are available in GenBank (http://www.ncbi.nlm.nih.gov). The ETS region was newly amplified here in *L. mariae-josephi* and other lupine species in order to gain more variable and informative characters from the nrDNA repeat (Badwin and Markos 1998; Bena et al. 1998). These regions were each successfully amplified in *L. mariae-josephi* and other lupine species (see Table 1 p. 179) following the PCR procedures previously described in Aïnouche and Bayer (1999) for the ITS region using primers ITS1 (5'-TCCGTAGGTGAACCTGCGG-3') and ITS4 (5'-TCCTCCGCTTATTGATATGC-3'); in Citerne (2005) for the *LEGCYC1A* region using primers LEGCYC1-F1 (5'-CTTCTACTTACAYWTCYTCAGGC-3') and LEGCYC1A-R1 (5'-CTACYACTACCCCTTCTGG-3'); and in Chandler et al. (2001) for the ETS region using the forward primer 18S-IGS (5′-CACATGCATGGCTTAATCTTTG-3′) and the reverse primer 281F (5'-TGCTTCCATTTGCTTGCTTGCCT-3′) designed by P. Cubas and C. Pardo (Universidad Complutense, Madrid, Spain; personal communication) to amplify a 100 bp longer ETS fragment.

PCR amplification of each of these regions was conducted following the same procedure previously described by Aïnouche and Bayer (1999) and Aïnouche et al. (2003). Each PCR reaction was performed in a volume of 50 µL containing 5 µL of 10X Taq-buffer (including 11 mM $MgCl_2$), 5 µL of a 2 mM mix of equimolar dNTPs, 0.5 µM of each primer (forward and reverse, according to the target region), 2.5 U of Red Taq DNA polymerase (Sigma), and 10-50 ng of DNA. The PCR amplification was started with 3 min of DNA denaturation at 94°C, prior to perform 30 cycles of 1 min denaturation at 94°C, 1 min at 48°C for primer annealing and 2 min of extension at 72°C for each cycle. A 7 min final extension at 72°C followed cycle 30. Double-stranded PCR products were purified with the EZ-10 Spin Column PCR Products Purification Kit (Bio Basic Inc.) and sent to Macrogen Inc. company (Seoul, South Korea) for direct sequencing using the ABI system. No significant sequence heterogeneity was previously detected among lupine samples sequenced for ITS, IGS and *rbcL* regions, or in this study for ETS sequences. However, regarding the ambiguous status of *L. mariae-josephi* and that a hybrid origin could not be excluded, purified ITS and ETS PCR products were cloned and sequenced in order to verify the sequence homogeneity. Cloning was performed using the pGEM-T easy Vector System and the Wizard Plus Minipreps kits, following the manufacturer's procedures (Promega), then positive plasmids containing an insert were sequenced employing T7 and SP6 primers. Thirteen and five clones of ITS and ETS regions, respectively, were obtained for *L. mariae-josephi*.

After PCR amplifications, double-stranded PCR products were purified with the EZ-10 Spin Column PCR Products Purification Kit (Bio Basic Inc.) and sent to Macrogen Inc. (Seoul, South Korea) for direct sequencing using the ABI system.

No significant sequence heterogeneity was previously detected among lupine samples sequenced for ITS and *LEGCYC1A* regions, from previous studies, or in this study for ETS sequences. However, regarding the ambiguous status of *L. mariae-josephi*, sequence homogeneity of purified ITS and ETS PCR products was verified by cloning, This was performed using the pGEM-T easy Vector System and the Wizard Plus Minipreps kits, following manufacturer's instructions (Promega). Positive plasmids containing an insert were sequenced employing T7 and SP6 primers.

Three separate data sets were constructed with ITS, ETS and *LEGCYC1A* sequences, obtained from GenBank or generated for this study, including almost all Mediterranean and African Old World lupines, and some representatives of main New World lupine clades, following insights from previous phylogenetic analyses of the genus (Käss and Wink 1997; Aïnouche and Bayer 1999; Aïnouche et al. 2003; Aïnouche et al. 2004; Hughes and Eastwood 2006; Drummond

2008). Here, *Lupinus affinis* J. Agardh, *L. argenteus* Pursh and *L. mexicanus* Cerv. ex. Lag. represent western American lupines; *L. villosus* Willd. represents unifoliolate lupines from southeastern United States (Florida); and *L. paraguariensis* Chodat & Hassler represents eastern South American species and their relatives from Texas (Aïnouche et al. 2004). Sequences from *Stauracanthus genistoides* (Brot.) Samp. and *Argyrolobium uniflorum* (Decne.) Jaub. & Spach, which are members of groups closely related to *Lupinus* (Aïnouche et al. 2003), were used as outgroups. For each data set, sequences were aligned with the software M-Coffee (Wallace et al. 2006; Moretti et al. 2007). Insertion-deletion events were coded with the software Seqstate (Müller 2005, 2006) using the multiple complex indel coding method (Simmons and Ochoterena 2000).

Phylogenetic analyses were conducted using PAUP* 4.0b10 (Swofford, 2003) and PhyML 3 (Guindon and Gascuel 2003). All data sets were subjected to Maximum Parsimony (MP) and Maximum Likelihood (ML) analyses. MP analyses were performed with heuristic searches and default options. For all data sets, ML analyses were performed following the procedure described in Harrison and Langdale (2006). The appropriate nucleotide substitution model of sequence evolution was determined by Modeltest 3.7 (Posada and Crandall 1998), using the Akaike information criterion (Posada and Buckley 2004). In each analysis, bootstrap method (1,000 replicates) was employed to estimate the robustness of clades (Felsenstein 1985). Following preliminary separate analyses, two additional ML analyses were conducted: one on a combined data set including both ITS and ETS sequences, and the other combining ITS, ETS and *LEGCYC1A* data sets (only including taxa for which sequences are available for all three DNA regions).

All taxa used in this study are presented in Table 1 p. 179, including their distribution and origins, their chromosome numbers, their sources and references, and their GenBank accession numbers.

Results and Discussion

General morphology and habitat—To compare *L. mariae-josephi* with other Old World lupine species, selected diagnostic morphological characters are reported in Table 2 p. 180, according to the data available from Gladstones (1974, 1984, 1998), Castroviejo and Pascual (1999) and Pascual (2004). Apart from its seed coat macro-texture, *L. mariae-josephi* is clearly distinct for most diagnostic features from all the Old World rough-seeded lupines (*Scabrispermae*), including *L. cosentinii*, the only one of them present in Spain. Compared to the Old World smooth seeded lupines, *L. mariae-josephi* shares more similarities in general morphology and habit with members of sections *Lutei* (*L. luteus*, *L. hispanicus*) and *Micranthi* (*L. micranthus*) than with *L. albus* and *L. angustifolius* (Table 2). Based on its short plant height, its leaf shape and pubescence, the number of leaflets per leaf, *L. mariae-josephi* resembles to *L. hispanicus* subsp. *bicolor* (*syn. L. gredensis*) and *L. micranthus* (but differs from these species for all other characters). Additionally, *L. mariae-josephi* exhibits marked differences with all Old World and American lupines occurring in the Iberian Peninsula (Castroviejo and Pascual 1999). These differences concern important diagnostic features, such as the size and structure of inflorescences and flowers, the flower color and the banner position (Table 2; Pascual 2004).

Besides, it is noteworthy from the ecological point of view that known populations of *L. mariae-josephi* only occur on calcareous basic soils, which is not in accordance with the common edaphic and pH range (acidic to neutral soils) generally tolerated by lupines (Howieson et al. 1998). Interestingly, only a few seeds of *L. mariae-josephi* germinated and grew when cultivated out of their native soil, and none of the young plants reached a fully developed stage and flowered in the greenhouse at Rennes (France), where all other Old World lupines normally do.

Seed-coat micomorphology—Different SEM micromorphological seed views of *L. mariae-josephi* are shown in Figure A.2 p. 181.

Also are presented for comparison the corresponding seed coat micromorphological views of samples from *L. luteus* and *L. princei*, which represent the typical patterns observed in the smooth seeded section *Lutei* and in the rough seeded group, respectively. Compared to the smooth surface of *L. luteus* (Fig. A.2b), the overall view of the seeds (x 10-12; Figs. A.2a, A.2b, A.2c) shows that *L. mariae-josephi* displays a rough seed surface, similar to that of the *Scabrispermae* (Fig. A.2c). However, *L. mariae-josephi* presents an ellipsoidal-orbicular flattened seed shape, which is rather different from the quadrangular shape with a prominent hilum observed in the *Scabrispermae* (Table 2; Fig. A.2c). As can be seen from Figs. A.2j, A.2k and A.2l, the transverse section of the lupine seed testa is constituted of three layers: an internal parenchymatous tissue, overlayed by two external unicellular layers, the inner sclerified hypodermis (composed of osteosclereid or hourglass cells with large intercellular spaces) and the outer layer of elongated palisade cells (also called macrosclereids or Malpighian cells). Macrosclereids are always longer than osteosclereids at midseed. In all the smooth seeded lupines, macrosclereids form a cohesive palisade layer and the micromorphological pattern observed at the external surface of the seed corresponds to the top of a single palisade cell, which is more or less prominent and sculptured according to the species. Previous studies have shown the high diagnostic value of this character for the Old World lupines (Heyn and Herrnstadt 1977; Aïnouche and Bayer 2000; Aïnouche et al. 2004). Each circum-Mediterranean smooth seeded species is well characterized by its cohesive palisade cells and its own unicellular pattern. The typical seed coat pattern of members of section *Lutei* is illustrated here by that of *L. luteus* (Figs. A.2e, A.2h). Instead, the rough seeded lupines (mostly African) all exhibited a similar and unique seed coat pattern well represented here by that of *L. princei* (Figs. A.2f, A.2i, A.2l). Compared to that of the smooth seeded lupines, this pattern is characterized by longer palisade cells, which are less cohesive and are distinctly fascicled in their upper part (Fig. A.2l), forming prominent pluricellular protuberances (or tubercles, separated by wide spaces) at the seed surface (Figs. A.2F, A.2I, A.2L). A distinct thin cuticular membrane stretched over and between tubercle tops contributes to the design of this typical pattern. Compared to these micromorphological types, *L. mariae-josephi* displays an intriguing intermediate seed coat pattern with on one side, a tuberculated seed surface (Figs. A.2d, A.2g), as for the *Scabrispermae*; and on the other side, an overall transverse section similar to that of smooth-seeded species (Fig. A.2j), but with prominent tubercles surmounting the external seed surface. Figure A.2j indicates that the palisade cell has not all the same height, and that only the highest among them contribute to form pluricellular prominent tubercles at the surface. Thus, this pattern shows at the same time similarities and differences with patterns observed in rough- and smooth-seeded Old World lupines. To date, none of the New World lupines surveyed exhibited a similar seed coat pattern.

Chromosome counts—Among dozens of cells observed, 21 intact metaphasic ones were selected for mitotic chromosome counts. Most of them showed more than 48 chromosomes, with an optimum number of $2n = 52$ well spread chromosomes, which is illustrated in Figure A.1. Among the Old World lupines, this number was only found in smooth seeded lupines: *L. luteus*, *L. hispanicus* subsp. *hispanicus*, *L. hispanicus* subsp. *bicolor* (syn. *L. gredensis*) and *L. micranthus*. Other numbers reported in this group were $2n = 50$ and 40 for *L. albus* and *L. angustifolius*, respectively; whereas rough seeded lupines have lower chromosome numbers, ranging from $2n = 32$ to 42 (Gladstones 1998; and references therein).

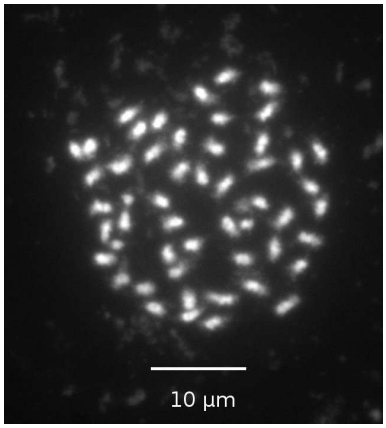

Figure A.1 — Somatic chromosome number of *Lupinus mariae-josephi* H. Pascual, $2n = 52$.

In the New World, the number $2n$ = 52 was reported for *L. villosus* and *L. cumulicola* (Conterato and Schifino-Wittmann 2006), two members of the singular unifoliolate monophyletic group from southeastern North America. Interestingly, recent molecular phylogenetic analyses supported the latter as a monophyletic group, which exhibited closer relationships to the Old World lupines than to the New World ones (Hughes and Eastwood 2006; Drummond 2008). Most frequent numbers recorded in the New World were $2n$ = 48 for almost all western American lupines (Dunn 1984; Conterato and Schifino-Wittmann 2006; and references therein); and $2n$ = 36 (with some putative dysploids exhibiting $2n$ = 32 or 34) for southeastern South American and southern North American (Texas) lupines (Turner 1957; Maciel and Schifino-Wittmann 2002; Conterato and Schifino-Wittmann 2006).

Phylogenetic position of L. mariae-josephi—ETS and *LEGCYC1A* sequences obtained for the four populations of *L. mariae-josephi* were identical. Only one nucleotide divergence was observed in one of the ITS sequences (0.2%). Three aligned sequences data matrices were constructed with ITS, ETS and *LEGCYC1A* sequences, each including representatives of the four *L. mariae-josephi* accessions. Each data matrix was subjected to maximum parsimony (MP) and maximum likelihood (ML) analyses to evaluate the phylogenetic position of *L. mariae-josephi*.

Used here for the first time to infer phylogenetic relationships within *Lupinus*, ETS sequences exhibited more parsimony informative characters (118 for 486 aligned nucleotide positions) than the commonly used ITS region (33 for 468 aligned nucleotide positions). This is in agreement with the ETS potential observed in various phylogenetic studies in Angiosperms, including Fabaceae (Badwin and Markos 1998; Bena et al. 1998; Chandler et al. 2001; Urbatsch et al. 2003; Suárez-Santiago et al. 2007). Separate MP or ML analyzes of ITS or ETS sequences resolved nearly the same groups, including the new lineage *L. mariae-josephi*, with however, a lack of resolution between main lineages (results not shown). To summarize information provided by both nrDNA spacers, a tree generated from the ML analysis of combined ITS and ETS sequences, using the TrN+G nucleotide substitution model (Tamura and Nei 1993), is presented in Figure A.3a. This phylogeny is consistent in several points with previous ITS phylogenies of *Lupinus* (Käss and Wink 1997; Aïnouche and Bayer 1999; Aïnouche et al. 2003; Aïnouche et al., 2004). As can be seen from Figure A.3a, main geographical lineages of lupines previously circumscribed are represented. In the New World, the *L. affinis–L. argenteus–L. mexicanus* group representing western American lupines receives a strong additional support from ETS data (99% of bootstrap support). *Lupinus villosus and L. paraguariensis*, representatives of southeastern North American unifoliolate lupines on one side, and of eastern South American lupines on the other side, are placed as two poorly supported sister lineages (63,7% of bootstrap), independent from the western New World clade. In the Old World, two lineages are highly supported: the rough seeded lupines (*Scabrispermae*) with 100% of bootstrap support; and the sister sections *Lutei* (incl. *L. luteus and L. hispanicus*) and *Angustifoli* (incl. *L. angustifolius*) with 97.5% bootstrap support. The other group, composed of the two Mediterranean smooth seeded lin-

eages *L. albus* and *L. micranthus*, is only poorly supported (65.1% of bootstrap). In this phylogeny, *L. mariae-josephi* clearly appears as a well-supported and distinct lineage (100% of bootstrap) within the weakly supported Old World assemblage. Its relationship with the group containing *L.albus*, *L. micranthus* and *Scabrispermae* is not supported by the bootstrap test (50.9%). Overall, relationships between main Old and New World lineages remain widely unresolved by ITS and ETS data.

The *LEGCYC1A* gene was previously shown to yield more phylogenetic signal than ITS, and to provide more well supported nodes in *Lupinus* (Ree et al. 2004; Hughes and Eastwood 2006). Here, *LEGCYC1A* sequences of *L. mariae-josephi* were compared to a reduced data set of orthologous sequences obtained from GenBank, and representative of major lupine lineages. The tree generated from the ML analysis, using the TIM3+G nucleotide substitution model (Posada and Buckley 2004), is shown in Figure A.3b. The topology obtained with this reduced data set is in general accordance with the topology obtained by Hughes and Eastwood (2006) with a broader sampling, and is illustrative of the new phylogenetic insights provided by the *LEGCYC1A* locus. Among these, the data supports the western and eastern New World lupines (excepted *L. villosus*), represented by *L. argenteus* and *L. paraguariensis*, respectively, as members of a well-supported monophyletic group (87.6% of bootstrap). The latter was not detected by ITS and ETS data. This New World clade is placed as sister to the assemblage composed of the Old World lineages and *L. villosus*. Within this assemblage, relationships among Old World lineages are poorly resolved and show some remarkable differences when compared to the ITS+ETS topology (Figure A.3a), such as for instance the decay of the sister relationship between *L. angustifolius* and *L. luteus–L. hispanicus*. Surprisingly, *L. mariae-josephi* rather shows a closer affinity with *L. villosus* (with a moderate bootstrap support of 76.3%) than to any other Old World lupines. This result was confirmed with a similar level of support by the analysis of a large data set including most *LEGCYC1A* sequences available in GeneBank (not shown). This relationship was also retrieved by another analysis of combined *LEGCYC1A* and ITS sequences (not shown). Recently, a phylogenetic analysis based on plastid data also provided some evidence supporting a close relationship between southeastern North American unifoliolate lupines and Old World lupines (Drummond 2008). Combining the three DNA regions ITS, ETS and *LEGCYC1A* (Fig. A.3c) provided more support to the New World lupines clade, exclusive of *L. villosus* (98.9% bootstrap); to the rough-seeded group (100% bootstrap); and re-establish a sister relationship between *L. angustifolius* and *L. luteus–L. hispanicus* (99.1% of bootstrap). However, this topology does not retain a close relationship between *L. villosus* and *L. mariae-josephi*. *Lupinus villosus* is placed as sister to New World lupines (with 75.1% of bootstrap), whereas the placement of *L. mariae-josephi* remains unresolved, indicating that the variable placement of *L. villosus* is mainly due to a conflict between ETS and *LEGCYC1A* data. The fact that this species possesses homogeneous ETS arrays of the highly repeated nrDNA cistron close to those of the New World lupines, and that it has an orthologous *LEGCYC1A* copy close to those of Old World lupines, suggests that *L. villosus* might have experienced particular evolutionary processes, such as for instance, reticulate evolution, concerted evolution among nrDNA repeats of different origins, and/or other forms of recombination leading to genetic heterogeneity and incongruence between gene trees (Seelanan et al. 1997; Wendel and Doyle 1998; Álvarez and Wendel 2003; Small et al. 2004; Soltis et al. 2008). However, this should be carefully examined using a broader taxonomic sampling of the ETS region, and employing other low copy and informative genes to avoid potential artifactual effects from unequal taxonomic weighting and representation, or from long-branch attraction/repulsion.

Finally, there is evidence from these re-

sults that the ETS nrDNA region introduced in this study provides additional support to some major *Lupinus* clades detected (as shown above) in previous phylogenies based on different nuclear and plastid sequences (including ITS and *LEGCYC1A*) and on a broader sampling (Käss and Wink 1997; Aïnouche and Bayer 1999; Aïnouche et al. 2004; Hughes and Eastwood 2006; Drummond 2008). Evidence from all these data still leaves uncertainties at the base of the genus, especially with regard to the placement and relationships of most Old World lineages, and of the singular unifoliolate southwestern North American lupines (represented by *L. villosus*). However, ETS data reveal a remarkable conflicting position of *L. villosus*, which was not detected before by either separate or combined analyses of ITS and *LEGCYC1A* sequences. Within this framework, *L. mariae-josephi* appears as another and intriguing distinct lupine lineage that remains unresolved within the Old World assemblage (with some unsuspected phylogenetic affinities with *L. villosus* according to *LEGCYC1A* data). Elucidation of the ETS–*LEGCYC1A* conflict regarding the latter species, which may results from a reticulate and/or a recombination process, will probably shed light on the placement of *L. mariae-josephi*.

Concluding remarks

When putting together all the observations and data presently available from various approaches, *L. mariae-josephi* appears as a distinct and enigmatic Old World lineage. Some of its similarities mentioned above with members of Old World lupines (in morphology and cytology), particularly *L. luteus*, *L. hispanicus* and *L. micranthus*, found no support from present molecular data. Also, it is noteworthy that *L. mariae-josephi* exhibits a rough seed coat surface type with an intermediate structure, which has never been seen before in *Lupinus*. However, data from *LEGCYC1A* gene provided some evidence suggesting phylogenetic affinities between *L. mariae-josephi* and *L. villosus*, representative of the singular unifoliolate southwestern North American lupines. These two species also share the same chromosome number, $2n = 52$.

Regarding the genetic and morphological peculiarities of *L. mariae-josephi*, it appears that phylogenetic placement of this recently discovered species highlights the complexity of the early evolutionary history of *Lupinus*. Therefore, it is obvious from this study that the accurate resolution of *L. mariae-josephi*'s placement is highly dependent on the clarification of the origins and relationships of Old World lineages and unifoliolate North American lupines, but also depends on the elucidation of the genetic heterogeneity of the latter. Further analyses including additional genes are required to answer these questions, which are central to the resolution of basal relationships of the genus to better understand the origin and the evolutionary processes that have accompanied the early diversification of the lupines.

Besides, the enigmatic *L. mariae-josephi* is of great interest from the ecological and agronomical perspective, as it is the first identified lupine that is restricted to basic calcareous soils. Most lupines, including those which have a high agronomic interest as nitrogen and protein suppliers (*L. albus, L. angustifolius, L. luteus, L. mutabillis*), are known to grow on sandy and sandy-loamy acid to neutral soils (Gladstones 1998; Howieson et al. 1998). That edaphic restriction represents one of the most important limits to the expansion of lupine cultivation in wide areas of poor calcareous soils in the Mediterranean region.

References

(see p. 191)

Table 1. List of *Lupinus* and outgroup taxa included in this study. Taxa are presented with their origin, geographic distribution, life history trait (LH: A, annual, or P, perennial), reference number, and GenBank accession numbers of their sequences analysed here.

Taxon	Origin/Distribution[a]	LH	Sample source[b]/ Reference number	Sequence accession numbers[c]		
				ITS1 / ITS2 or Complete ITS region	ETS	LEGCYC1A
Argyrolobium uniflorum	?/OW, Afr	P	- /G25	Z95566/Z95567	**XX**	–
Stauracanthus genistoides	Spain/Afr-Med	P	MAF 7908/P72	AF384340/AF384341	**XX**	DQ529865
Lupinus affinis	Oregon/NW, NA	A	USDA504315/N20	AF007487	**XX**	–
Lupinus albus	Algeria/OW, Med	A	INAE-DZ/M20	AH006088	**XX**	DQ529787
Lupinus angustifolius subsp. *angustifolius*	Algeria/OW, Med	A	AKA-M1/T24	Z72202/Z72203	**XX**	DQ529757
Lupinus argenteus	Washington/NW, NA	P	USDA504374/N23b	AF007458	**XX**	AY338910
Lupinus atlanticus	Morocco/OW, Afr-Med	A	USDA384612/T2	AH006080	**XX**	–
Lupinus cosentinii	?/OW, Med	A	INRAL-FR/T15	**GU058032**	**XX**	DQ529744
Lupinus digitatus	Egypt/OW, Afr-Med	A	WADA-PI26877/K34	AY338948	**XX**	AY338922
Lupinus hispanicus	Portugal/OW, Med	A	USDA384555/T22	AH006096	**XX**	DQ529780
Lupinus luteus	Algeria/OW, Med	A	AKA-M5/T20	AF007478	**XX**	DQ529782
Lupinus mariae-josephi	Spain/OW, Med	A	H. Pascual/MJ1	**GU058033**	**XX**	**GU045296**
Lupinus mariae-josephi	Spain/OW, Med	A	CIEF/V39Q	**GU058034**	**XX**	**GU045297**
Lupinus mariae-josephi	Spain/OW, Med	A	CIEF/V39T	**GU058035**	**XX**	**GU045298**
Lupinus mariae-josephi	Spain/OW, Med	A	CIEF/V39V	**GU058036**	**XX**	**GU045299**
Lupinus mexicanus	Mexico, F.D/NW, CA	P	USDA14748/N51	AH006086	**XX**	–
Lupinus micranthus	Algeria/OW, Med	A	AKA-M8/T28	AF007480	**XX**	DQ529813
Lupinus paraguariensis	S.C., Brazil/NW, SA	P	BRA-02828/BZ1	AF007476	**XX**	DQ529928
Lupinus pilosus	?/OW, Med-Afr	A	INAE-DZA13/T6	AH006081	**XX**	–
Lupinus villosus	Florida/NW, SE-NA	A	D. Jones/K36	AY609189/AY609192	**XX**	DQ529749

Abbreviation correspondence: [a] OW, Old World; NW, New World; NA, North America; SA, South America; CA, Central America; Afr, Africa; Med, Mediterranean.
[b] MAF, Herbario Facultad Farmacia, Madrid; USDA, US Department of Agriculture, Washington; INAE-DZ, Institut National d'Agronomie, El-Harrach, Algérie; AKA, Abdelkader Aïnouche; INRAL, INRA, Lusignan, France; BRA, EMBRAPA, Brasil.
[c] Sequence accession numbers in bold were generated from this study; all other sequences were obtained from GenBank databases.

Table 2. Selected diagnostic morphological characters (from Gladstones, 1974, 1984, 1998; Castroviejo and Pascual, 1999; Pascual, 2004) used to compare *Lupinus mariae-josephi* to representatives of most lupine taxa occurring in the Mediterranean region and North Africa.

	L. mariae-josephi	*L. micranthus*	*L. luteus*	*L. hispanicus*	*L. albus*	*L. angustifolius*	*L. cosentinii*	*L. digitatus*	*L. pilosus*	*L. atlanticus*
Plant height (cm)	**<30**	10–50	20–80	<70	30–120	20–150	20–120	15–40	30–80	<60
No. of leaflets per leaf	**5–7**	5–7	7–9	4–9	5–9	5–9	9–11	9–11	9–11	9–11
leaflets size (mm)	32 × 8	10–50 × 6–20	30–60 × 8–15	15–50 × 5–10	20–60 × 12–20	15–35 × 1.5–4	25–60 × 7–12	20–50 × 5–10	25–60 × 10–18	—
Leaflets shape	Oblong-Lanceolate	Obovate cuneate/oblong	Obovate-oblong	Obovate-oblong	Obovate-oblong	Linear, narrow	Oblong-oblanceolate	Oblong-obovate	Oblong-obovate	Oblong_oblanceolate
Leaflets pubescence above/ below	**Glabrous/hairy**	Villous / villous	Villous / sparsely	±glabrous / sparsely	±glabrous / hairy	Glabrous / sparsely	Sericeous / sericeous	Sparcely sericeous/ densely sericeous	Densely sericeous/ densely sericeous	Sericeous / sericeous
Inflorescence	**Subverticillate**	Subverticillate	Verticillate	Verticillate	Subverticillate	Subverticillate	Verticillate	Verticillate	Subverticillate-Verticillate	Subverticillate-Verticillate
Inflorescence size (mm	**38**	30–120	50–250	50–250	50–300	50–200	50–150	30–150	100–300	—
Flower pedoncule (mm)	3–4	2–3	1–2	1–2	1–2	2–4	2–4	—	Long	—
Flower color	**white-yellowish or redish**	Blue	Bright golden yellow	Bluish, lilac or pinkish	White, tinged blue or violet	Blue tinged purple	Blue+yellowish spot on standart	Blue+white or yellow spot	Blue+central white spot or band on standart	Blue +white or yellowish central sector on standart
Flower length x height (mm)	12 × 6	Varying greatly	14–16 × 14–16	-	15–16 × 12–14	11–15 × 10–14	12–17 × 14–19	16 × 18	15–20 × 16–22	18 × 20–22
Pod length x width (mm)	50 × 20	22–38 × 14–16	40–60 × 10-14	40–60 × 6-12	70–150 × 12–20	35–50 × 7–10	40–55 × 13–16	30–60 × 9–12	50–80 × 20–25	40–70 × 15–22
Seeds per pod	**3–4**	2–5	4–6	4–6	3–6	4–7	3–5	3–4	2–4	3–5
Seed shape	Elipsoidal flattened	Lensy- flattened	Orbicular-quadrangular	Orbicular-quadrangular	Square, compressed	±globular	Orbicular-quadrangular	Orbicular-quadrangular	Orbicular weakly flattened	Oblong compressed
Seed size (mm)	8 × 6 × 4	5–8 × 4–6 × 3–4	6–9 × 5–8 × 3-4	4–8	8–14 × 6–12 × 2–5	4–6 × 3–5 × 3–4	6–9 × 4–7 × 3–4	7 × 6 × 3	10–14 × 9–12 × 6-8	8–11 × 6–8 × 4–5
Seed coat surface	**Rough**	Smooth	Smooth	Smooth	Smooth	Smooth	Tuberculate	Rough	Rough	Rough

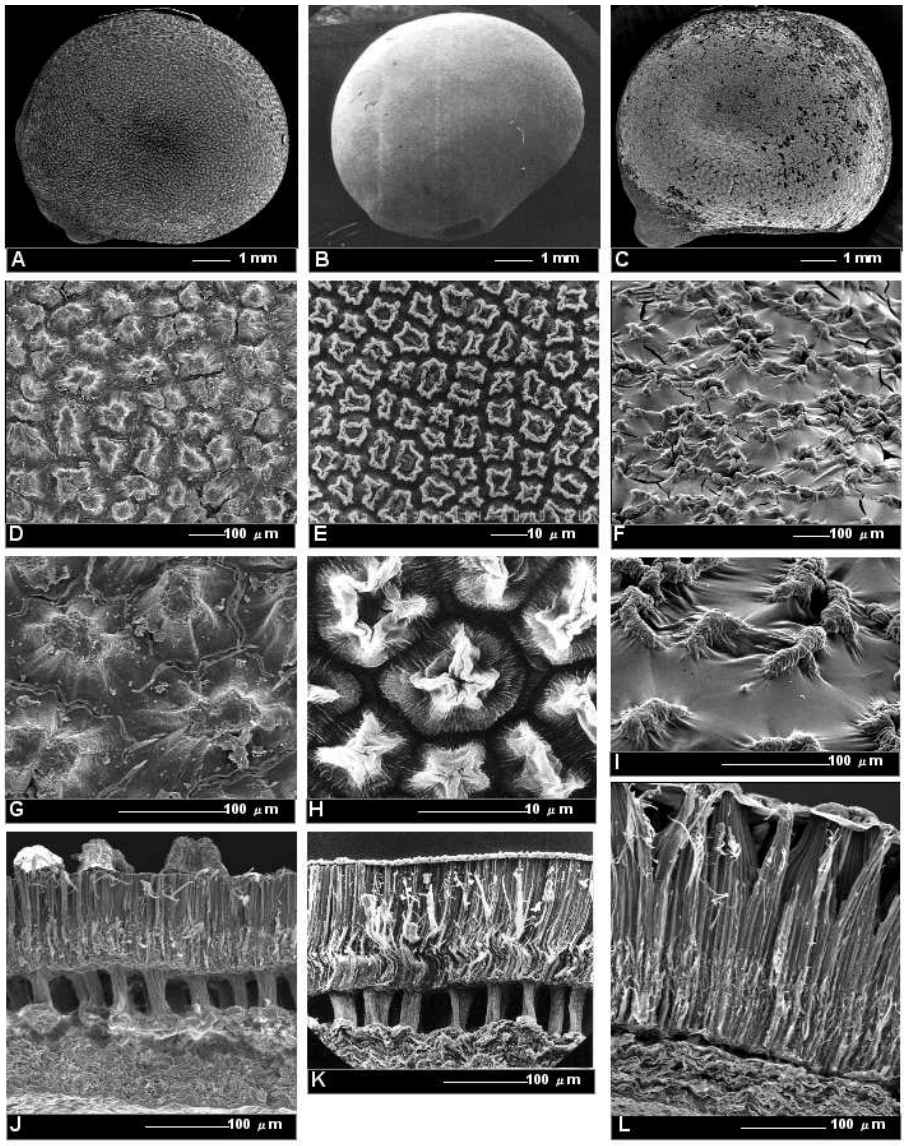

FIG. A.2— Seed coat micromorphological patterns of three Old World lupines. A-D-G-J: *Lupinus mariae-josephi*; B-E-H-K: *L. luteus*; C-F-I-L: *L. princei*. A-B-C: Overview of seeds; D-E-F and G-H-I: mid-seed surface patterns at low and high magnificence; J-K-L: transverse section of the testa. Small letters on the SEM micrographs indicate: c, cuticular membrane; m, macrosclereids or palisade cells; o, osteosclereids or hypodermis cells; p, internal parenchymatous tissue; t, tubercle or fascicled palisade cells (pluricellular pattern); u, unicellular pattern.

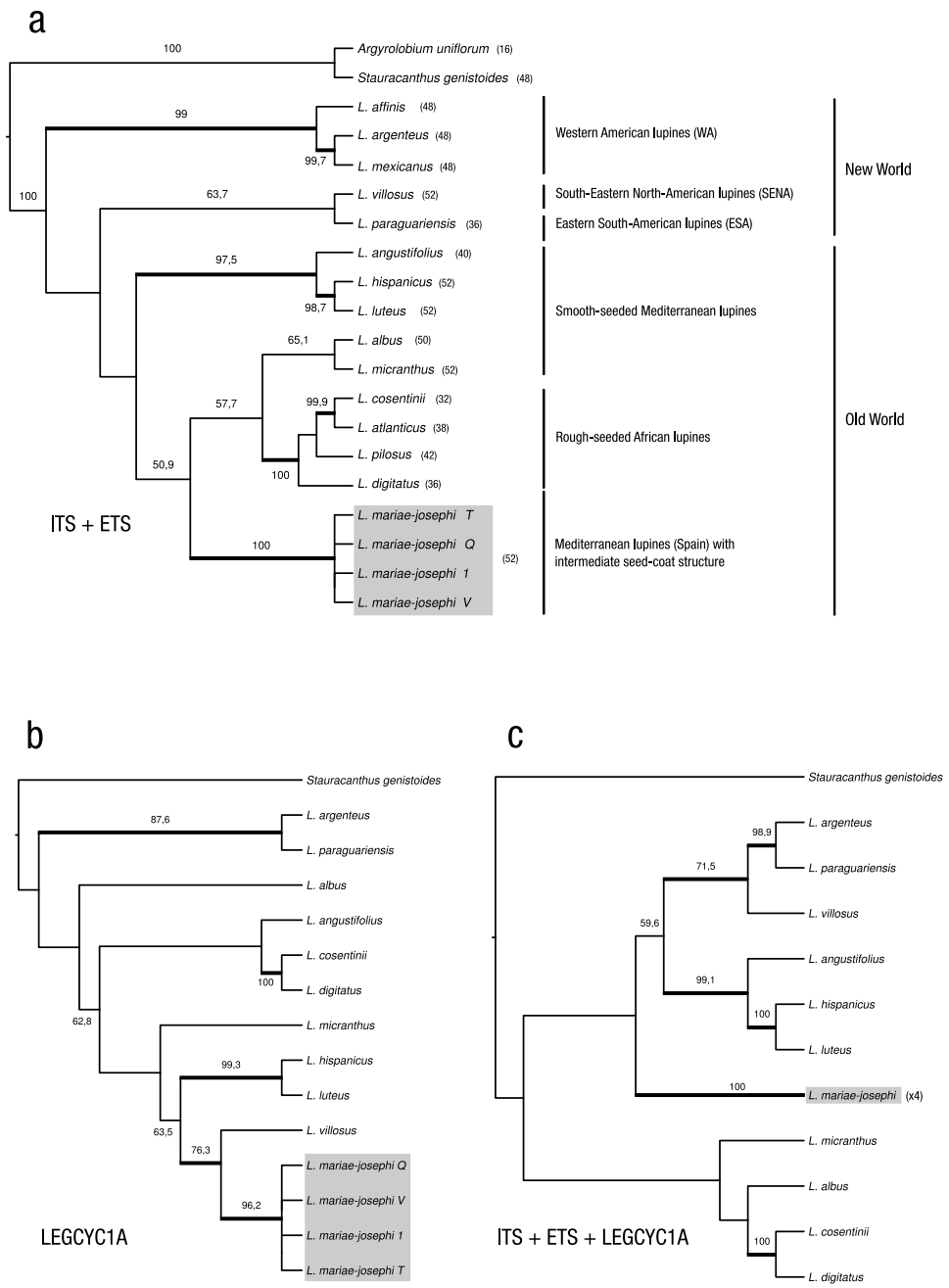

FIG. A.3 — Variable phylogenetic placement of *Lupinus mariae-josephi* according to the analyzed nuclear DNA region. The phylogenetic trees presented were obtained through maximum likelihood, with 1,000 non-parametric bootstrap replicates. Bootstrap support is placed on the branches for values superior or equal to 50%, with bold branches highlighting support equal or superior to 70%. Chromosome numbers follow taxon names. a) The first tree is generated from a combination of nrRNA data (ITS and ETS) for 18 lupines and 2 other genistoids. Remarkable geographic and taxonomic groups of lupines previously identified by earlier phylogenetic analysis are noted on the right (Aïnouche et al. 2004; Hughes and Eastwood 2006). b) Tree generated from a dataset of 11 LEGCYC1A sequences (from GenBank), complemented with 4 sequences from different populations of *L. mariae-josephi*. c) Tree generated from the combination of three DNA regions ITS, ETS and LEGCYC1A.

B

Code génétique

> « Le chemin qui mena de la découverte de la structure en double hélice de l'ADN au décryptage du code génétique est tellement tortueux qu'une présentation chronologique en serait totalement confuse. Pour résoudre le problème du rôle des gènes dans la synthèse des protéines, deux approches étaient possibles : dans l'une, d'inspiration génétique, les chercheurs tentaient de déduire de la structure de l'ADN (et de l'action des gènes dans la cellule) le rôle de cette molécule dans la synthèse des protéines. Dans l'autre, plus biochimique, ils essayaient de reconstituer des systèmes acellulaires capables de synthétiser les protéines. C'est cette dernière voie de recherche qui permit le décryptage du code génétique. »
>
> — Michel Morange (1994, p. 156)

> « Lorsque le déchiffrage du code génétique fut entrepris, le lien entre ADN, protéines et ARN, et le rôle des différents ARN dans la synthèse protéique restait encore à éclaircir. Il fallait pour cela que soit découvert l'ARN messager. »
>
> — Michel Morange (1994, p. 179)

Le décryptage du code génétique pris plusieurs années, de 1953 à 1961, et passa par l'élaboration de plusieurs modèles et de nombreuses hypothèses erronées, jusqu'à l'expérience de Johann Matthaei et Marshall Nirenberg (Matthaei & Nirenberg, 1961 ; Nirenberg & Matthaei, 1961). La communication de leurs résultats, lors du v^{e} congrès international de biochimie à Moscou en août 1961, eu un effet important sur la communauté. Les résultats s'accumulèrent rapidement, et en 1966, la correspondance entre les triplets (puisqu'il s'agit d'un code à trois lettres) et les acides aminés était établie, le déchiffrage était complet.

code IUPAC	signification
A	Adénine
C	Cytosine
G	Guanine
T	Thymine
U	Uracile
R	G ou A (puRine)
Y	T ou C (pYrimidine)
K	G ou T (Keto)
M	A ou C (aMino)
S	G ou C (Strong)
W	A ou T (Weak)
B	C, G ou T (non A)
D	A, G ou T (non C)
H	A, C ou T (non G)
V	A, C ou G (non T)
N	A, C, G ou T (aNy)

5′	2^e position				3′
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

acide aminé	code 3 lettres	code 1 lettre
alanine	Ala	A
arginine	Arg	R
asparagine	Asn	N
acide aspartique	Asp	D
cystéine	Cys	C
acide glutamique	Glu	E
glutamine	Gln	Q
glycine	Gly	G
histidine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
méthionine	Met	M
phénylalanine	Phe	F
proline	Pro	P
sérine	Ser	S
thréonine	Thr	T
tryptophane	Trp	W
tyrosine	Tyr	Y
valine	Val	V
tous	–	Z

C

Liste des amorces

Toutes les amorces sont données dans le sens 5′ vers 3′.

ETS

- Gast1 : CGG TTG CGG CTC TGG TGT TC ;
- 18S-IGS : CAC ATG CAT GGC TTA ATC TTT G ;
- Lu1-Gast1 : CGG TTG CGG CTC TGG TGT TCA AT ;
- Ge1-Gast1 : CGG TTG CGG CTC TGG TGT TCA CTC G ;
- 281F : TGC TTC CAT TTG CTT GCT TGC CT.

Région *trn*L-*trn*F

- C : CGA AAT CGG TAG ACG CTA CG ;
- D : GGG GAT AGA GGG ACT TGA AC ;
- E : GGT TCA AGT CCC TCT ATC CC ;
- F : ATT TGA ACT GGT GAC ACG AG ;

ITS

- ITS1 : TCC GTA GGT GAA CCT GCG G ;
- ITS4 : TCC TCC GCT TAT TGA TAT GC ;

***rbc*L**

- N : ATG TCA CCA CAA ACA GAA ACT AAA GC ;
- R : TAT CCA TTG CTG GGA ATT CAA ATT TG ;

SymRK

- F0 : TCA TCA GAT CAG CTT CTG CAA ;
- F1 : CTG CAA CTG AAG GGT TTG AGA GCA ;

- F3 : CAA TAG ATC ATA GCT GGT TCT CTG ;
- F5 : CTT AAT CTA ACC TTG GTC AAG GC ;
- F5B[1] : CCT CGA GTA ATC TCA AAG GAA C ;
- F5C : GTC CTA CAA ACA GCT CTT ACT ;
- F6 : GGG GTC ACT TCC AGA ATC AA ;
- F7 : GTG GAT CAA TCT TGA TTA CTT TGG C ;
- R1 : CCA CCT CTA TGT ATT CCA AAG TG ;
- R3 : GCC AAA GTA ATC AAG ATT GAT CCA C ;
- R3a : TTG ATT CTG GAA GTG ACC CC ;
- R3P : CAA GCA GTG CTT GTT GAT CTG TC ;
- R4A : GKA TCA CTT CCA CTG AWG AAA TG ;
- R5 : GCC TTG ACC AAG GTT AGA TTA AG ;
- R5c : AGT AAG AGC TGT TTG TAG GAC ;
- R7 : GTC CTG GAG GCT GGA TGA TA ;
- R8 : CAG AGA ACC AGC TAT GAT CTA TTG ;

LEGCYC1A

- F1 : CTT CTA CTT ACA YWT CYT CAG GC ;
- A-R1 : CTA CYA CTA CCC CTT CTG G ;

Ty1/*copia*

- U : ACN GCN TTY YTN CAY GG ;
- D : ARC ATR TCR TCN ACR TA ;

Ty3/*gypsy*

- GyRT1 : MRN ATG TGY GTN GAY TAY MG ;
- GyRT4 : RCA YTT NSW NAR YTT NGC R ;

Plasmide pGEM-T

- T7 : TAA TAC GAC TCA CTA TAG GG ;
- SP6 : ATT TAG GTG ACA CTA TAG ;

1. Certain tubes d'amorces ont été étiquetés par erreur « F5P ».

D

Liste des taxa

Le tableau présenté page suivante regroupe la liste des taxa utilisés au cours de ce travail de thèse. Notre échantillonnage couvre la quasi-totalité des lupins de l'Ancien Monde et inclu également des lupins aux statuts incertains comme *Lupinus mariae-josephi* ou *L. pilosus tassilicus*. Pour ce qui est du Nouveau Monde, nous avons pris soin d'inclure des représentants des principaux groupes de lupins décrits dans la littérature.

TAB. D–1 — Liste des taxa utilisés au cours de ce travail de thèse. Feuilles simples (S), digitées (D) ou mixtes (M). Cycle de vie annuel (A) ou pérenne (P). Données tirées de Turner (1957) et Planchuelo (1994) pour les lupins de l'Ouest du Nouveau Monde, Dunn (1971) ; Maciel & Schifino-Wittmann (2002) ; Conterato & Schifino-Wittmann (2006) pour les lupins unifoliolés du Nouveau Monde ; Gladstones (1974, 1998), Plitmann & Pazy (1984), Obermayer *et al.* (1999), Pascual (2004) et Fos Martín *et al.* (2006) pour les lupins de l'Ancien Monde.

Taxon	Code	Source[1]	Origine	Distribution	2*n*	Feuilles	Cycle
Ancien Monde							
L. albus L.	M20	INRAL-FR/M20	Grèce	Méditerranée	50	D	A
——— var. *graecus* (Boiss. & Sprun.) Franko & Silva	M12	INRAL-FR/M12	Grèce	Méditerranée	50	D	A
L. anatolicus W. Święcicki & W. K. Święcicki	K32	AKA/ANAT-SWIEC.	Turquie	Afrique	42	D	A
L. angustifolius L. ssp. *angustifolius*	T24	AKA/ANT37	Tizi-ouzon, Algérie	Méditerranée	40	D	A
——— ssp. *reticulatus* Desv.	T25	AKA/ANGH0	Île d'Hœdic, France	Méditerranée	40	D	A
L. atlanticus Gladstones	T1	USDA/384612-FM83-T1	Maroc	endémique	38	D	A
———	T2	USDA/384613-FM87-T2	Maroc	endémique	38	D	A
———	T11	INRA-SAPF	Maroc	endémique	38	D	A
L. cosentinii Guss.	T7	INRAL-FR/A16	France	Afrique	32	D	A
L. digitatus Forsskål	T4	WADA/PI 26877	Égypte	Afrique	36	D	A
L. hispanicus spp. *bicolor* (Merino) Gladstones	T23	USDA/PI 384554	Espagne	Méditerranée	52	D	A
——— spp. *hispanicus* Boiss. & Reut.	T22	USDA/PI 244461	Portugal	Méditerranée	52	D	A
L. luteus L.	M5	AKA/M5	Zeralda, Algérie	Méditerranée	52	D	A
L. micranthus Guss.	T19	AKA/M8	Algérie	Méditerranée	52	D	A
L. mariae-josephi H. Pascual	MJ1	Madrid/711480	Espagne	endémique	52 (?)	D	A
———	MJT	CIEF V39Q	Espagne	endémique	52 (?)	D	A
———	MJQ	CIEF V39T	Espagne	endémique	52 (?)	D	A
———	MJV	CIEF V39V	Espagne	endémique	52 (?)	D	A
L. palaestinus Boiss.	T14	INRAL-FR/A15	Proche-orient	Méditerranée-Afrique	42	D	A
L. pilosus Murray	A13	INAE-DZ/A13	Tizi-ouzon, Algérie	Afrique	42	D	A
———	T10	Institut Vavilov/T10	Afrique du Nord	Afrique	42	D	A
———	T13	USDA W6 PI 11995/T13	Afrique du Nord	Afrique	42	D	A
——— ssp. *tassilicus*	A641	Hb.LeHouerou/LI 1404641	Lybie	Afrique	?	D	A
L. princei Harms	T0	WADA P23021/T0	Kenya	Afrique	38	D	A
———	T16	RP Chyulu 1800/T16	Chyulu Hills, Kenya	Afrique	38	D	A
———	T17	RP Chyulu 1915/T17	Chyulu Hills, Kenya	Afrique	38	D	A
L. vavilovi (Atab.) Kurl. & Stankev.[2]	VAV1	Institut Vavilov/K3118	Yougoslavie	Méditerranée	50	D	A

2. *Lupinus vavilovi* (Atabekova & Maissurjan) Kurl. & Stankev. est un synonyme de la sous-espèce *L. albus* L. var. *graecus* (Przyborowski & Weeden, 2001), elle même souvent désignée par *L. graecus* (Naganowska *et al.*, 2003).

Liste des taxa utilisés au cours de ce travail de thèse

Taxon	Code	Source	Origine	Distribution	2*n*	Feuilles	Cycle
Nouveau Monde							
L. albicaulis Douglas ex. Hook.	N27[3]	USDA/284705	California, USA	Ouest de l'Am. du Nord	?	D	P
L. albifrons Benth. ex Lindl.	?	USDA/284707	California, USA	Ouest de l'Am. du Nord	?	D	P
L. alopecuroides Desr.	Q3	M.-Th. Misset	Quilochoa, Équateur	Ouest de l'Am. du Sud	?	D	P
	S7	16 579	Pérou	Ouest de l'Am. du Sud	?	D	P
L. affinis J. Agardh	N20	USDA/504315	Oregon, USA	Ouest de l'Am. du Nord	48	D	A
L. arboreus Sims	N88	USDA/393932	Nouvelle-Zélande	Ouest de l'Am. du Nord	48	D	P
L. arcticus S. Wats.	?	Hb. ALTA/95826	Yukon, Canada	Nord-Ouest de l'Am. du Nord	48	D	P
L. aridus Dougl.	?	?	?	Ouest de l'Am. du Nord	48	D	P
L. arizonicus S. Wats.	N81	?	?	Sud-ouest de l'Am. du Nord	48	D	A
L. argenteus Pursh	N23	USDA/504374	Washington, USA	Ouest de l'Am. du Nord	48	D	P
L. bracteolaris Desr.	S80	USDA/404349	Brésil	Sud-Est de l'Am. du Sud	32-34	D	A
L. cacuminus Standl.	K39	?	Mexico F.D.	Amérique centrale	?	D	?
L. campestris Cham. & Schltdl.	K38	?	Mexico F.D.	Amérique centrale	?	D	A
L. concinnus J. Agardh	N19	USDA/284715	Californie, USA	Ouest de l'Am. du Nord	48	D	A
L. densiflorus Benth. var. *densiflorus*	?	USDA/15617	?	Ouest de l'Am. du Nord	48	D	A
L. diffusus Nutt.	K35			Est de l'Am. du Nord		S	P
L. duranii Eastwood	?	Hb ALTA/92238	California, USA	Sud-ouest de l'Am. du Nord	?	?	P
L. elegans H.B.K.	S33	USDA/185099	Mexico F. D.	Amérique centrale	48	D	P
L. excubitus M.E. Jones	?	Hb ALTA/95550	California, USA	Sud-ouest de l'Am. du Nord	?	D	P
L. gibertianus C.P. Smith	1835	UFRGSB/1835MTSW	Lages, SC, Brésil	Est de l'Am. du Sud	36	D	A
L. hirsutissimus Benth.	N85	Hb USDA/284719	California, USA	Sud-ouest de l'Am. du Nord	48	D	A
L. jaimehintoniana B.L. Turner	K22	?	Mexico F.D.	Amérique centrale	?	D	P
L. lanatus Benth.	?	?	?	Est de l'Am. du Sud	36	?	?
L. latifolius Lindl. ex J.G. Agardh	?	USDA/284720	California, USA	Ouest de l'Am. du Nord	?	D	P
L. lepidus Dougl. ex Lindl.	H6b	Hb ALTA/94855	Wyoming, USA	Nord-ouest de l'Am. du Nord	48	D	P
L. leucophyllus Dougl. ex Lindl.	?	USDA/504316	Oregon, USA	Ouest de l'Am. du Nord	48	D	P
L. littoralis Dougl.	?	USDA/504401	Washington, USA	Ouest de l'Am. du Nord	48	D	P
L. luteolus Kellogg	LT	USDA/284721	Californie, USA	Ouest de l'Am. du Nord	?	D	A
L. magnistipulatus Planchuelo & Dunn	1840	UFRGSB/1840MTSW	São Francisco de Paulo, RS, Brésil	Est de l'Am. du Sud	36	D	P
L. mexicanus Cerv. in Lag.	N51	USDA/14748	Mexico F.D.	Amérique centrale	48	D	P
L. minimus Dougl. ex Hook.	?	USDA/504439	Oregon, USA	Nord-ouest de l'Am. du Nord	?	D	P
L. microcarpus ?	T27	?	?	?	?	D	P
L. montanus ?	K37	?	Mexique	Amérique centrale	?	D	P
L. multiflorus Desr. in Lam.	S32	USDA/508613	Brésil	Est de l'Am. du Sud	36	D	P

3. Aussi étiqueté N90.

Liste des taxa utilisés au cours de ce travail de thèse

Taxon	Code	Source	Origine	Distribution	2*n*	Feuilles	Cycle
L. mutabilis Sweet	MU23	INAE-DZ/S35	Pérou	Ouest de l'Am. du Sud	48	D	A
L. nanus Dougl. ex Benth. 1835	N42	USDA/284729	Californie, USA	Ouest de l'Am. du Nord	48	D	A
L. nootkatensis Donn. ex Sims.	?	Ch. Biteau	Islande	Ouest de l'Am. du Nord	48	D	P
L. paraguariensis Chodat & Hassler	BZ1	CENARGEN/BRA-02828	Brésil	Est de l'Am. du Sud	36	M	P
	BZ3	?	Brésil	Est de l'Am. du Sud	36	M	P
L. polyphyllus Lindl.	T26	USDA/504404	Washington, USA	Nord de l'Am. du Nord	48	D	P
	N21b	?	Washington, USA	Nord de l'Am. du Nord	48	D	P
L. pusillus Pursh	?	USDA/504356	Oregon, USA	Ouest de l'Am. du Nord	48	D	A
L. sericeus Pursh	?	USDA/356830	Utah, USA	Ouest de l'Am. du Nord	48	D	P
L. sparsiflorus Benth.	N83	USDA/577289	Arizona, USA	Sud-ouest de l'Am. du Nord	48	D	A
L. subvexus C.P. Smith var. *subvexus*	?	?	?	Sud-ouest de l'Am. du Nord	48	D	A
L. succulentus Dougl. ex K. Koch	?	USDA/284728	California, USA	Sud-ouest de l'Am. du Nord	48	D	A
L. sulphureus Dougl. ex Hook.	?	USDA/504367	Washington, USA	Nord-ouest de l'Am. du Nord	48	D	P
L. texensis Hook.	N45	USDA/577291	Texas, USA	Sud de l'Am. du Nord	36	D	A
L. villosus Willd.	K36	?	Floride, USA	Est de l'Am. du Nord	52	S	A
Extra-groupes							
Anartrophyllum cummingii (Hooket Arn.) F. Phil.	201	?	?	?	?	–	?
Argyrolobium uniflorum (D.C) Jauber & Spach	G25	?	?	Afrique	16	–	?
Chamaecystus mollis (Cav.) Greuter & Burdet	C84	RBG-Kew/84327	?	Ancien Monde	?	?	P
Crotalaria podocarpa D.C	K50	RBG-Kew/90928	?	Ancien Monde	16	–	A
Crotalaria saharae Cosson	K51	RN 16-07	?	Ancien Monde	?	–	P
Cytisus heterochrous Colmeiro	G8	?	?	?	?	–	?
Genista tinctoria L.	G56	RBG-Kew/51334	?	Ancien Monde	48	–	P
Stauracanthus genistoides (Brot.) Samp.	P72	MAF[4]/7908	Helva, Espagne	Ancien Monde	48	–	P
Thermopsis rhombifolia (Pursh) Richardson var. *ovata*	G46	?	Idaho, USA	Nouveau Monde	18	–	P
Ulex australis Clemente	D27	?	?	?	?	–	P
Ulex europaeus L.	P91	?	?	?	?	–	P
Ulex parviflorus Pourr.	G24	?	?	?	?	–	P

4. Facultad de Farmacia, Universidad Complutense, Madrid

Bibliographie & références

Les références sont triées par ordre alphabétique suivant le nom de famille des auteurs, puis par date de publication, du plus ancien au plus récent.

Références

Abascal, F., Zardoya, R. & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, *21*(9), 2104-2105.

Abbasi, A. A. (2008). Are we degenerate tetraploids? More genomes, new facts. *Biology Direct*, *3*(1), 50.

Abd-Alla, M. H. (1998). Effect of *Lupinus* seed diffusates on *Bradyrhizobium* sp. growth and nodulation of lupine. *Folia Microbiologica*, *43*(2), 182-186.

Adams, K. L. & Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, *8*(2), 135-141.

Adhikari, K., Thomas, G., Buirchell, B. J. & Sweetingham, M. (2008). Identification of anthracnose resistance in yellow lupins and its incorporation into breeding lines. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 251-254). Fremantle, Australia.

Aïnouche, A. K. (1998). *Diversité et évolution du genre* Lupinus *L. (Fabaceae)*. Thèse de doctorat, université de Rennes 1. (N° d'ordre 1970, 135 p.)

Aïnouche, A. K. & Bayer, R. J. (1996). Phylogenetic relationships among and within the Old and New World *Lupinus* species (Fabaceae) based on Internal Transcribed Spacer sequences of nuclear ribosomal DNA. In G. D. Hill (Ed.), *Proceedings of the Eighth International Lupin Conference* (p. 236-244). Asilomar, California.

Aïnouche, A. K. & Bayer, R. J. (1999). Phylogenetic relationships in *Lupinus* (Fabaceae: Papilionoideae) based on Internal Transcribed Spacer sequences (ITS) of nuclear ribosomal DNA. *American Journal of Botany*, *86*(4), 590-607.

Aïnouche, A. K. & Bayer, R. J. (2000). Genetic evidence supports the new Anato-

lian lupine accession, *Lupinus anatolicus*, as an Old World "rough-seeded" lupine (section Scabrispermae) related to *L. pilosus*. *Folia Geobotanica*, *35*(1), 83-95.

Aïnouche, A. K., Bayer, R. J., Cubas, P. & Misset, M.-Th. (2003). Phylogenetic relationships within tribe Genisteae (Papilionaceae) with special reference to the genus *Ulex*. In B. B. Klitgaard & A. Bruneau (Eds.), *Advances in Legume Systematics. Part 10. Higher Level Systematics* (p. 239-252). Royal Botanic Gardens, Kew.

Aïnouche, A. K., Bayer, R. J. & Misset, M.-Th. (2004). Molecular phylogeny, diversification and character evolution in *Lupinus* (Fabaceae) with special attention to Mediterranean and African lupines. *Plant Systematics and Evolution*, *246*(3-4), 211-222.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Alix, K. & Heslop-Harrison, J. S. (Pat). (2004). The diversity of retroelements in diploid and allotetraploid *Brassica* species. *Plant Molecular Biology*, *54*(6), 895-909.

Allawi, H. T. & SantaLucia, J., Jr. (1997). Thermodynamics and NMR of internal G·T mismatches in DNA. *Biochemistry*, *36*(34), 10581-10594.

Allawi, H. T. & SantaLucia, J., Jr. (1998a). Nearest neighbor thermodynamic parameters for internal G·A mismatches in DNA. *Biochemistry*, *37*(8), 2170-2179.

Allawi, H. T. & SantaLucia, J., Jr. (1998b). Nearest-neighbor thermodynamics of internal A·C mismatches in DNA: sequence dependence and pH effects. *Biochemistry*, *37*(26), 9435-9444.

Allawi, H. T. & SantaLucia, J., Jr. (1998c). Thermodynamics of internal C·T mismatches in DNA. *Nucleic Acids Research*, *26*(11), 2694-2701.

Allen, J. E. & Salzberg, S. L. (2005). JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, *21*(18), 3596-3603.

Altschul, S. F., Gish, W. R., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403-410.

Álvarez, I. & Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, *29*(3), 417-434.

Andersson, I. & Backlund, A. (2008). Structure and function of Rubisco. *Plant Physiology and Biochemistry*, *46*(3), 275-291. (PMID: 18294858)

Andrews, C. B. & Gregory, T. R. (2009). Genome size is inversely correlated with relative brain size in parrots and cockatoos. *Genome*, *52*(3), 261-267.

Angiosperm Phylogeny Group II. (2003). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants. *Botanical Journal of the Linnean Society*, *141*(4), 399-436.

Anisimova, M. & Gascuel, O. (2006). Approximate Likelihood-Ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology*, *55*(4), 539-552.

Anxolabéhère, D., Nouaud, D., Quesneville, H. & Ronsseray, S. (2007). Transposons : des gènes anarchistes ? *Pour la Science*, *351*, 82-89.

Apostolico, A. & Denas, O. (2008). Fast algorithms for computing sequence distances by exhaustive substring composition. *Algorithms for Molecular Biology*, *3*(1), 13.

Arkhipova, I. R., Pyatkov, K. I., Meselson, M. & Evgen'ev, M. B. (2003). Retroelements containing introns in diverse invertebrate taxa. *Nature Genetics*, *33*(2), 123-124.

Arnoldi, A. (2008). Nutraceutical properties of white and narrow-leafed lupin. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 452-454). Fremantle, Australia.

Aslanidis, C. & Jong, P. J. de. (1990). Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Research*, *18*(20), 6069-6074.

Baer, E. von. (2008). Efficiency and quality in the production of sweet lupin. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 71-74). Fremantle, Australia.

Baetcke, K. P., Sparrow, A. H., Nauman, C. H. & Schwemmer, S. S. (1967). The relationship of DNA content to nuclear and chromosome volumes and to radiosensitivity (LD_{50}). *Proceedings of the National Academy of Sciences of the United States of America*, *58*(2), 533-540.

Bailey, C. D., Carr, T. G., Harris, S. A. & Hughes, C. E. (2003). Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. *Molecular Phylogenetics and Evolution*, *29*(3), 435–455.

Baldwin, B. G. & Markos, S. (1998). Phylogenetic utility of the External Transcribed Spacer (ETS) of 18S–26S rDNA: congruence of ETS and ITS trees of *Calycadenia* (Compositae). *Molecular Phylogenetics and Evolution*, *10*(3), 449-463.

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S. & Donoghue, M. J. (1995). The ITS region of nuclear ribosomal DNA: a valuable source of evidence on Angiosperm phylogeny. *Annals of the Missouri Botanical Garden*, *82*(2), 247-277.

Barac, T., Taghavi, S., Borremans, B., Provoost, A., Oeyen, L., Colpaert, J. V. *et al.* (2004). Engineered endophytic bacteria improve phytoremediation of water-soluble, volatile, organic pollutants. *Nature Biotechnology*, *22*(5), 583-588.

Barakat, A., Carels, N. & Bernardi, G. (1997). The distribution of genes in the genomes of Gramineae. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(13), 6857-6861.

Baranyi, M. & Greilhuber, J. (1995). Flow cytometric analysis of genome size variation in cultivated and wild *Pisum sativum* (Fabaceae). *Plant Systematics and Evolution*, *194*(3-4), 231-239.

Baranyi, M. & Greilhuber, J. (1996). Flow cytometric and Feulgen densitometric analysis of genome size variation in *Pisum*. *Theoretical and Applied Genetics*, *92*(3-4), 297-307.

Barriel, V. (1994). Phylogénies moléculaires et insertions–délétions de nucléotides. *Comptes Rendus de l'Académie des Sciences – Séries III – Sciences de la Vie*, *317*, 693-701.

Bartoš, J., Paux, É., Kofler, R., Havránková, M., Kopecký, D., Suchánková, P. *et al.* (2008). A first survey of the rye (*Secale cereale*) genome composition through BAC end sequencing of the short arm of chromosome 1R. *BMC Plant Biology*, *8*(1), 95.

Batzoglou, S. (2005). The many faces of sequence alignment. *Briefings in Bioinformatics*, *6*(1), 6-22.

Baum, D. A., Yoon, H.-S. & Oldham, R. L. (2005). Molecular evolution of the transcrip-

tion factor LEAFY in Brassicaceae. *Molecular Phylogenetics and Evolution*, *37*(1), 1-14.

Beaulieu, J. M., Leitch, I. J., Patel, S., Pendharkar, A. & Knight, Ch. A. (2008). Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytologist*, *179*(4), 975-986.

Beaulieu, J. M., Moles, A. T., Leitch, I. J., Bennett, M. D., Dickie, J. B. & Knight, Ch. A. (2007). Correlated evolution of genome size and seed mass. *New Phytologist*, *173*(2), 422-437.

Becker, G., Bianchi, H., Carrière, Ch., Coste, J.-P., Duffert, P. J., Dughi, R. *et al.* (1957). *Tournefort*. Muséum national d'histoire naturelle, Paris. (321 p.)

Bekaert, M. & Teeling, E. C. (2008). UniPrime: a workflow-based platform for improved universal primer design. *Nucleic Acids Research*, *36*(10), e56.

Bekpen, C., Marques-Bonet, T., Alkan, C., Antonacci, F., Leogrande, M. B., Ventura, M. *et al.* (2009). Death and resurrection of the human *IRGM* gene. *PLoS Genetics*, *5*(3), e1000403.

Bena, G., Jubier, M.-F., Olivieri, I. & Lejeune, B. (1998). Ribosomal External and Internal Transcribed Spacers: combined use in the phylogenetic analysis of *Medicago* (Leguminosae). *Journal of Molecular Evolution*, *46*(3), 299-306.

Bender, J. (1998). Cytosine methylation of repeated sequences in eukaryotes: the role of DNA pairing. *Trends in Biochemical Sciences*, *23*(7), 252-256.

Bennett, M. D. (1972). Nuclear DNA content and minimum generation time in herbaceous plants. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *181*(63), 109-135.

Bennett, M. D., Bhandol, P. & Leitch, I. J. (2000). Nuclear DNA amounts in angiosperms and their modern uses — 807 new estimates. *Annals of Botany*, *86*(4), 859-909.

Bennett, M. D. & Leitch, I. J. (2004). Plant DNA C-values database (release 5.0, Dec. 2004). http://www.rbgkew.org.uk/cval/homepage.html.

Bennett, M. D. & Leitch, I. J. (2005a). Genome size evolution in plants. In T. R. Gregory (Ed.), *The Evolution of the Genome* (p. 89-162). San Diego, California: Elsevier.

Bennett, M. D. & Leitch, I. J. (2005b). Plant genome size research: a field in focus. *Annals of Botany*, *95*(1), 1-6.

Bennett, M. D., Leitch, I. J. & Hanson, L. (1998). DNA amounts in two samples of Angiosperm weeds. *Annals of Botany*, *82*(6), 121-134.

Bennett, M. D., Price, H. J. & Johnston, J. S. (2008). Anthocyanin inhibits propidium iodide DNA fluorescence in *Euphorbia pulcherrima*: implications for genome size variation and flow cytometry. *Annals of Botany*, *101*(6), 777-790.

Bennetzen, J. L. (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, *115*(1), 29-36.

Bennetzen, J. L. & Kellogg, E. A. (1997). Do plants have a one-way ticket to genomic obesity? *Plant Cell*, *9*(9), 1509-1514.

Bennetzen, J. L., Ma, J. & Devos, K. M. (2005). Mechanisms of recent genome size variation in flowering plants. *Annals of Botany*, *95*(1), 127-132.

Benson, D. A., Boguski, M. S., Lipman, D. J. & Ostell, J. (1994). GenBank. *Nucleic Acids Research*, *22*(17), 3441-3444.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. (2009). GenBank. *Nucleic Acids Research*, *37*(suppl. 1), D26-31.

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*(2), 163-193.

Berk, A., Bramm, A., Böhm, H., Aulrich, K. & Rühl, G. (2008). The nutritive value of lupins in sole cropping systems and mixed intercropping with spring cereals for grain production. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 66-70). Fremantle, Australia.

Bestor, T. H. (2003). Cytosine methylation mediates sexual conflict. *Trends in Genetics*, *19*(4), 185-190.

Bhattacharyya, M. K., Smith, A. M., Ellis, T. H. N., Hedley, C. & Martin, C. (1990). The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. *Cell*, *60*(1), 115-122.

Biderre, C., Pagès, M., Méténier, G., Canning, E. U. & Vivarès, C. P. (1995). Evidence for the smallest nuclear genome (2.9 mb) in the microsporidium *Encephalitozoon cuniculi*. *Molecular and Biochemical Parasitology*, *74*(2), 229-231.

Biémont, Ch. & Vieira, C. (2006). Junk DNA as an evolutionary force. *Nature*, *443*(7111), 521-524.

Bishop, J. G. & Schemske, D. W. (1998). Variation in flowering phenology and its consequences for lupines colonizing Mount St. Helens. *Ecology*, *79*(2), 534-546.

Biteau, Ch. (2004). *Phylogénie et évolution de la taille des génomes chez les lupins (*Lupinus *L.).* Rapport de DEA éco-éthologie évolutive, université de Rennes 1.

Boersma, J. G., Pallotta, M., Li, C., Buirchell, B. J., Sivasithamparam, K. & Yang, H. (2005). construction of a genetic linkage map using MFLP and identification of molecular markers linked to domestication genes in narrow-leafed lupin (*Lupinus angustifolius* L.). *Cellular & Molecular Biology Letters*, *10*(2), 331-344.

Bofkin, L. & Goldman, N. (2007). Variation in evolutionary processes at different codon positions. *Molecular Biology and Evolution*, *24*(2), 513-521.

Bonnivard, É. & Higuet, D. (2009). Fluidity of eukaryotic genomes. *Comptes Rendus Biologies*, *332*(2-3), 234-240.

Bonvallot, V. (2004). Des plantes au service de la dépollution. *Biofutur*, *23*(242), 29-31.

Boschin, G., D'Agostina, A., Annicchiarico, P. & Arnoldi, A. (2008). Genotypic and genotype-environmental effects on fatty acid composition of *Lupinus albus* L. seed. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 321-323). Fremantle, Australia.

Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, *422*(6930), 433-438.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791-799.

Bragg, L. H. (1983). Seed coats of some *Lupinus* species. *Scanning Electron Microscope*, *4*, 1739-1745.

Brandt, J., Schrauth, S., Veith, A.-M., Froschauer, A., Haneke, T., Schultheis, Ch. *et al.* (2005). Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene*, *345*(1), 101-111.

Brennecke, S., Becker, W.-M., Lepp, U. & Jappe, U. (2007). Anaphylactic reaction to lupine flour. *Journal der Deutschen Dermatologischen Gesellschaft*, *5*(9), 774-776.

Briand, O. (2006). Pesticides et environnement. In Le Monde (Ed.), *Sur les chemins de la découverte* (p. 25-39). Presses Universitaires de France.

Britten, R. J. & Kohne, D. E. (1968). Repeated Sequences in DNA. *Science*, *161*(3841), 529-540.

Bucher, M., Wegmüller, S. & Drissner, D. (2009). Chasing the structures of small molecules in arbuscular mycorrhizal signaling. *Current Opinion in Plant Biology*, *12*(4), 500-507.

Buirchell, B. J. (2008). Narrow-leafed lupin breeding in Australia — Where to from here? In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 226-230). Fremantle, Australia.

Burge, Ch. B. & Karlin, S. (1997, apr). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, *268*(1), 78-94.

Burge, Ch. B. & Karlin, S. (1998). Finding the genes in genomic DNA. *Current Opinion in Structural Biology*, *8*(3), 346-354.

Camillo, M. F., Pozzobon, M. T. & Schifino-Witmann, M. T. (2006). Chromosome numbers in South American Andean species of *Lupinus* (Leguminosae). *Bonplandia*, *15*(3-4), 113-119.

Candole, A. P. de. (1813). *Théorie élémentaire de la botanique*. Deterville, Paris.

Cantrell, M. A., Scott, L., Brown, C. J., Martinez, A. R. & Wichman, H. A. (2008). Loss of LINE-1 activity in the megabats. *Genetics*, *178*(1), 393-404.

Cao, Y., Adachi, J., Janke, A., Pääbo, S. & Hasegawa, M. (1994). Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. *Journal of Molecular Evolution*, *39*(5), 519-527.

Capoen, W., Goormachtig, S., De Rycke, R., Schroeyers, K. & Holsters, M. (2005). SrSymRK, a plant receptor essential for symbiosome formation. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10369-10374.

Capy, P. (2005). Classification and nomenclature of retrotransposable elements. *Cytogenetic and Genome Research*, *110*(1-4), 457-461.

Carstairs, S. A., Buirchell, B. J. & Cowling, W. A. (1992). Chromosome number, size and intercrossing ability of three Old World lupines, *Lupinus princei* Harms, *L. atlanticus* Gladstones and *L. digitatus* Forskål, and implications for cyto-systematic relationships among the rough-seeded lupins. *Journal of the Royal Society of Western Australia*, *75*, 83-88.

Casacuberta, J. M. & Santiago, N. (2003). Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, *311*, 1-11.

Casola, C., Hucks, D. & Feschotte, C. (2008). Convergent domestication of pogo-like

transposases into centromere-binding proteins in fission yeast and mammals. *Molecular Biology and Evolution*, *25*(1), 29-41.

Castro Guerra, E. de. (2007). *Inkscape*. CampusPress.

Castroviejo, S. & Pascual, H. (1999). *Flora iberica — Plantas vasculares de la Península Ibérica e Islas Baleares. Vol. VII(I). Leguminosae* (S. Talavera, C. Aedo, S. Castroviejo, L. Romero Zarco C. Sáez, F. J. Salgueiro & M. Velayos, Eds.). Real Jardín Botánico, CSIC, Madrid.

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *Journal of Cell Science*, *34*(1), 247-278.

Cavalier-Smith, T. (1985). Cell volume and the evolution of eukaryotic genome size. In T. Cavalier-Smith (Ed.), *The Evolution of genome size* (p. 104-184). John Wiley & Sons, Chichester, United Kingdom.

Cavalier-Smith, T. (2005). Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Annals of Botany*, *95*(1), 147-175.

Cavalli-Sforza, L. L. & Edwards, A. W. F. (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, *19*(3, Part I), 233-257.

Chandler, G. T., Bayer, R. J. & Crisp, M. D. (2001). A molecular phylogeny of the endemic Australian genus *Gastrolobium* (Fabaceae: Mirbelieae) and allied genera using chloroplast and nuclear markers. *American Journal of Botany*, *88*(9), 1675-1687.

Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B. *et al.* (2005). Molecular basis of evolutionary events that shaped the *Hardness* (*Ha*) locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell*, *17*(4), 1033-1045.

Chase, M. W., Soltis, D. E., Omstead, R. G., Morgan, D., Les, D. H., Mishler, B. D. *et al.* (1993). Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. *Annals of the Missouri Botanical Garden*, *80*(3), 528-580.

Chen, M.-J., Shimada, T., Moulton, A. D., Harrison, M. & Nienhuis, A. W. (1982). Intronless human dihydrofolate reductase genes are derived from processed RNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(23), 7435-7439.

Cheng, X., Zhang, D., Cheng, Z., Keller, B. & Ling, H.-Q. (2009). A new family of Ty1-copia-like retrotransposon originated in the tomato genomes by a recent horizontal transfer event. *Genetics*, *181*(4), 1183-1193.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. *et al.* (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, *31*(13), 3497–3500.

Choisne, N., Demange, N., Orjeda, G., Michelet, L., Pelletier, E., Salanoubat, M. *et al.* (2004). Le séquençage de génomes de plantes. In *La génomique en biologie végétale* (p. 33-57). INRA éditions.

Chou, H.-H. & Holmes, M. H. (2001). DNA sequence quality trimming and vector removal. *Bioinformatics*, *17*(12), 1093-1104.

Chudy, M., Leśniewska, K., Wolko, B. & Święcicki, W. K. (2008). Narrow-leafed lupin (*Lupinus angustifolius* L.) comparative studies. In J. A. Palta & J. B. Berger (Eds.),

Proceedings of the 12th International Lupin Conference — Lupins for health and wealth (p. 332-335). Fremantle, Australia.

Church, K. W. & Helfman, J. I. (1993). Dotplot: a program for exploring self-similarity in millions of lines of text and code. *Journal of Computational and Graphical Statistics*, *2*(2), 153-174.

Citerne, H. L. (2005). A primer set for specific amplification of two cycloidea-like genes in the genistoid clade of *Leguminosae* subfam. *Papilionoideae*. *Edinburgh Journal of Botany*, *62*(3), 119-126.

Citerne, H. L., Luo, D., Pennington, R. T., Coen, E. & Cronk, Q. C. (2003). A phylogenomic investigation of CYCLOIDEA-Like TCP genes in the Leguminosae. *Plant Physiology*, *131*(3), 1042-1053.

Clarke, L. & Carbon, J. (1976). A colony bank containing synthetic CoI EI hybrid plasmids representative of the entire *E. coli* genome. *Cell*, *9*(1), 91-99.

Clauset, A., Shalizi, C. R. & Newman, M. E. J. (2007). Power-law distributions in empirical data. arXiv:0706.1062v2 [physics.data-an]. (43 p., version 2 publiée en février 2009, `http://arxiv.org/abs/0706.1062`)

Clements, J. C., Buirchell, B. J. & Cowling, W. A. (1996). Relationship between morphological variation and geographical origin or selection history in *Lupinus pilosus*. *Plant Breeding*, *115*(1), 16-22.

Clements, J. C., Sweetingham, M. S., Smith, L., Francis, G., Thomas, G. & Sipas, S. (2008). Crop improvement in *Lupinus mutabilis* for australian agriculture — Progress and prospects. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 244-250). Fremantle, Australia.

Commission Directive 2006/142/EC of 22 December 2006 amending Annex IIIa of Directive 2000/13/EC of the European Parliament and of the Council listing the ingredients which must under all circumstances appear on the labelling of foodstuffs. (2006). *Official Journal — European Union Legislation*, *49*(368), 110-111.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674-3676.

Conterato, I. F. & Schifino-Wittmann, M. T. (2006). New chromosome numbers, meiotic behaviour and pollen fertility in American taxa of *Lupinus* (Leguminosae): contributions to taxonomic and evolutionary studies. *Botanical Journal of the Linnean Society*, *150*(2), 229-240.

Cordaux, R., Udit, S., Batzer, M. A. & Feschotte, C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(21), 8101-8106.

Cotton, J. A. & Wilkinson, M. (2009). Supertrees join the mainstream of phylogenetics. *Trends in Ecology & Evolution*, *24*(1), 1-3.

Cowling, W. A., Buirchell, B. J. & Tapia, M. E. (1998). *Lupin.* Lupinus *L.* (n° 23). Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany/International Plant Genetic Resources Institute, Rome, Italy. (105 p.)

Cremonini, R., Colonna, N., Stefani, A., Galasso, I. & Pignone, D. (1994). Nuclear DNA

content, chromatin organization and chromosome banding in brown and yellow seeds of *Dasypyrum villosum* (L.) P. Candargy. *Heredity*, *72*(4), 365-373.

Crisp, M. D., Gilmore, S. & Wyk, B.-E. van. (2003). Molecular phylogeny of the genistoid tribes of papilionoid legumes. In P. S. Herendeen & A. Bruneau (Eds.), *Advances in Legume Systematics. Part 9* (p. 249-276). Royal Botanic Gardens, Kew.

Cristofolini, G. (1989). A serological contribution to the systematics of the genus *Lupinus* (Fabaceae). *Plant Systematics and Evolution*, *166*(3-4), 265-278.

Cronk, Q. C. B. (2006). Legume flowers bear fruit. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(13), 4801-4802.

Cubas, P., Lauter, N., Doebley, J. & Coen, E. (1999). The TCP domain: a motif found in proteins regulating plant growth and development. *Plant Journal*, *18*(2), 215-222.

Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics*, *14*(10), 892-893.

Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J. *et al.* (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research*, *16*(6), 738-749.

Dadejová, M., Lim, K. Y., Soucková-Skalická, K., Matyášek, R., Grandbastien, M.-A., Leitch, A. R. *et al.* (2007). Transcription activity of rRNA genes correlates with a tendency towards intergenomic homogenization in *Nicotiana* allotetraploids. *New Phytologist*, *174*(3), 658-668.

Dalevi, D., Desantis, T. Z., Fredslund, J., Andersen, G. L., Markowitz, V. M. & Hugenholtz, Ph. (2007). Automated group assignment in large phylogenetic trees using GRUNT: GRouping, Ungrouping, Naming Tool. *BMC Bioinformatics*, *8*(1), 402.

Dante Alighieri. (v. 1308-1321). *Divina Commedia*. (Traduit de l'italien par Antoine de Rivarol, *La Divine comédie*. Disponible sur `http://fr.wikisource.org/wiki/La_Divine_Com%C3%A9die`)

Darlu, P. & Tassy, P. (1993). *La Reconstruction phylogénétique. Concepts et méthodes*. Masson. (Cet ouvrage est disponible en ligne grâce au travail de Yann Bertrand et Régis Debruyne : `http://lis.snv.jussieu.fr/sfs/publications_sfs.shtml`)

Darwin, Ch. R. (1859). *The Origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* (6^{e} éd.). London: John Murray. (Traduit de l'anglais par Edmond Barbier, *L'Origine des espèces au moyen de la sélection naturelle ou la lutte pour l'existence dans la nature*, 1876. Révisé par Daniel Becquemont, *Garnier-Flammarion*, Paris, 1982)

Dashti, N., Khanafer, M., El-Nemr, I., Sorkhoh, N., Ali, N. & Radwan, S. (2009). The potential of oil-utilizing bacterial consortia associated with legume root nodules for cleaning oily soils. *Chemosphere*, *74*(10), 1354-1359.

Daskalos, A., Nikolaidis, G., Xinarianos, G., Savvari, P., Cassidy, A., Zakopoulou, R. *et al.* (2008). Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. *International Journal of Cancer*, *124*(1), 81-87.

Davis, C. C., Webb, C. O., Wurdack, K. J., Jaramillo, C. A. & Donoghue, M. J. (2005). Explosive radiation of Malpighiales supports a mid-cretaceous origin of modern tropical rain forests. *The American Naturalist*, *165*(3), E36-E65.

De Bodt, S., Maere, S. & Peer, Y. Van de. (2005). Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*, *20*(11), 591-597.

De Mita, S., Santoni, S., Ronfort, J. & Bataillon, Th. (2007). Adaptive evolution of the symbiotic gene NORK is not correlated with shifts of rhizobial specificity in the genus *Medicago*. *BMC Evolutionary Biology*, *7*, 210.

Demmel, A., Hupfer, C., Ilg Hampe, E., Busch, U. & Engel, K.-H. (2008). Development of a Real-Time PCR for the detection of lupine DNA (*Lupinus* species) in foods. *Journal of Agricultural and Food Chemistry*.

Derelle, É., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A. Z., Robbens, S. *et al.* (2006). Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(31), 11647-11652.

d'Errico, F. & Sánchez Goñi, M. F. (2003). Neandertal extinction and the millennial scale climatic variability of OIS 3. *Quaternary Science Reviews*, *22*(8-9), 769-788.

Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Research*, *12*(7), 1075-1079.

Diao, X., Freeling, M. & Lisch, D. (2006). Horizontal transfer of a plant transposon. *PLoS Biology*, *4*(1), e5.

Dijkink, B., Miedendorp de Bie, V. & Blom, W. (2008). Altering lupin flour for the food industry. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 455-458). Fremantle, Australia.

Dmitriev, D. A. & Rakitov, R. A. (2008). Decoding of superimposed traces produced by direct sequencing of heterozygous indels. *PLoS Computational Biology*, *4*(7), e1000113.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M. & Batzoglou, S. (2005). ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research*, *15*(2), 330-340.

Doležel, J. & Bartoš, J. (2005). Plant DNA flow cytometry and estimation of nuclear genome size. *Annals of Botany*, *95*(1), 99-110.

Doolittle, W. F. & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, *284*(5757), 601-603.

Douady, Ch. J., Catzeflis, F., Raman, J., Springer, M. S. & Stanhope, M. J. (2003). The Sahara as a vicariant agent, and the role of Miocene climatic events, in the diversification of the mammalian order Macroscelidea (elephant shrews). *Proceedings of the National Academy of Sciences of the United States of America*, *100*(14), 8325-8330.

Doyle, J. J. & Luckow, M. A. (2003). The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiology*, *131*(3), 900-910.

Drummond, A. J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214.

Drummond, Ch. S. (2008). Diversification of *Lupinus* (Leguminosae) in the western New World: Derived evolution of perennial life history and colonization of montane habitats. *Molecular Phylogenetics and Evolution*, *48*(2), 408-421.

Drummond, Ch. S. & Hamilton, M. B. (2007). Hierarchical components of genetic variation at a species boundary: population structure in two sympatric varieties of *Lupinus microcarpus* (Leguminosae). *Molecular Ecology*, *16*(4), 753-769.

Dunn, D. B. (1971). A case of long range dispersal and "rapid speciation" in *Lupinus*. *Transactions of the Missouri Academy of Science*, *5*, 26-38.

Dunn, D. B. (1984). Cytotaxonomy and distribution of the New World lupin species. In *Proceedings of the Third International Lupin Conference* (p. 67-85). La Rochelle, France.

Dunn, D. B. & Gillett, J. M. (1966). *The Lupines of Canada and Alaska*. Queen's Printer, Ottawa, Canada. (89 p.)

Dunn, D. B. & Harmon, W. E. (1977). The *Lupinus montanus* complex of Mexico and Central America. *Annals of the Missouri Botanical Garden*, *64*(2), 340-356.

Dupressoir, A., Vernochet, C., Bawa, O., Harper, F., Pierron, G., Opolon, P. *et al.* (2009). Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(29), 12127-12132.

Dwivedi, B. & Gadagkar, S. R. (2009). Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evolutionary Biology*, *9*(1), 211.

Eastwood, R. J., Drummond, Ch. S., Schifino-Wittmann, M. T. & Hughes, C. E. (2008). Diversity and evolutionary history of lupins — Insights from new phylogenies. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 346-354). Fremantle, Australia.

Eastwood, R. J. & Hughes, C. E. (2008). Origins of domestification of *Lupinus mutabilis* in the Andes. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 373-379). Fremantle, Australia.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113.

Edgar, R. C. & Batzoglou, S. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology*, *16*(3), 368-373.

Edwards, A. C. & Barneveld, R. J. van. (1998). Lupins for livestock and fish. In J. S. Gladstones, C. Atkins & J. Hamblin (Eds.), *Lupins as crop plants: biology, production and utilization* (p. 385-411). Wallingford, UK: CAB International.

Egan, A. N. & Crandall, K. A. (2008). Divergence and diversification in north american psoraleeae (fabaceae) due to climate change. *BMC Biology*, *6*(1), 55.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, *8*(3), 163-167.

Endre, G., Kereszt, A., Kevei, Z., Mihacea, S., Kaló, P. & Kiss, G. B. (2002). A receptor kinase gene regulating symbiotic nodule development. *Nature*, *417*(6892), 962-966.

Evans, G. M., Rees, H., Snell, C. L. & Sun, S. (1972). The relationship between nuclear DNA amount and the duration of the mitotic cycle. *Chromosomes Today*, *3*, 24-31.

Ewing, B. & Green, P. (1998). Base-Calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, *8*(3), 186-194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated

sequencer traces using phred. I. Accuracy assessment. *Genome Research*, *8*(3), 175-185.

Fæstea, C. K., Løvikb, M., Wikerb, H. G. & Egaasa, E. (2004). A case of peanut cross-allergy to lupine flour in a hot dog bread. *International Archives of Allergy and Immunology*, *135*(1), 36-39.

Farrell, B. D. (1998). "Inordinate Fondness" explained: why are there so many beetles? *Science*, *281*(5376), 555-559.

Fawcett, J. A., Maere, S. & Peer, Y. Van de. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(14), 5737-5742.

Felsenstein, J. (1978a). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, *27*(4), 401-410.

Felsenstein, J. (1978b). The number of evolutionary trees. *Systematic Zoology*, *27*(1), 27-33.

Felsenstein, J. (1985a). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, *39*(4), 783-791.

Felsenstein, J. (1985b). Confidence limits on phylogenies with a molecular clock. *Systematic Zoology*, *34*(2), 152-161.

Felsenstein, J. (2001). The troubled growth of statistical phylogenetics. *Systematic Biology*, *50*(4), 465-467.

Fernandes, A. & Queiros, M. (1978). Contribution à la connaissance cytotaxonomique des spermatophyta du Portugal. *Boletim da Sociedade Broteriana*, *IV*(suppl. 2), 79-164.

Ferragina, P., Giancarlo, R., Greco, V., Manzini, G. & Valiente, G. (2007). Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. *BMC Bioinformatics*, *8*, 252.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397-405.

Filatov, D. A., Howell, E. C., Groutides, C. & Armstrong, S. J. (2009). Recent spread of a retrotransposon in the *Silene latifolia* genome, apart from the Y chromosome. *Genetics*, *181*(2), 811-817.

Filipowicz, W., Jaskiewicz, L., Kolb, F. A. & Pillai, R. S. (2005). Post-transcriptional gene silencing by siRNAs and miRNAs. *Current Opinion in Structural Biology*, *15*(3), 331-341.

Findley, R. (1981). Mont St. Helens. The day the sky fell. *National Geographic*, *159*(1), 50-65.

Finn, R. N. & Kristoffersen, B. A. (2007). Vertebrate vitellogenin gene duplication in relation to the "3R Hypothesis": correlation to the pelagic egg and the oceanic radiation of teleosts. *PLoS ONE*, *2*(1), e169.

Flavell, A. J., Smith, D. B. & Kumar, A. (1992). Extreme heterogeneity of *Ty1-copia* group retrotransposons in plants. *Molecular and General Genetics*, *231*(2), 233-242.

Flavell, R. B. (1994). Inactivation of gene expression in plants as a consequence of

specific sequence duplication. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(9), 3490-3496.

Flavell, R. B., Rimpau, J. & Smith, D. B. (1977). Repeated sequences DNA relationships in four cereal genomes. *Chromosoma*, *63*(3), 205-222.

Folkman, W., Szerechan, J. & Gulewicz, K. (2002). Preparations of alkaloid-rich lupin in plant protection: an effect of the preparations on feeding and development of *Pieris brassicae* L. and *Pieris rapae* L. *Journal of Plant Protection Research*, *42*(2), 143-155.

Fontaine, A., Monte, A. de & Touzet, H. (2008). MAGNOLIA: multiple alignment of protein-coding and structural RNA sequences. *Nucleic Acids Research*, *36*(suppl. 2), W14-18.

Fortuné, Ph. M., Roulin, A. & Panaud, O. (2008). Horizontal transfer of transposable elements in plants. *Communicative & Integrative Biology*, *1*(1), 74-77.

Fos Martín, S., Navarro Peris, A., Ferrando Pardo, I., Alba Villegas, S. & Laguna Lumbreras, E. (2006). Nuevas poblaciones del altramuz valenciano (*Lupinus mariae-josephi*). *Toll Negre*, *8*, 21-26.

Francis, D., Davies, M. S. & Barlow, P. W. (2008). A strong nucleotypic effect on the cell cycle regardless of ploidy level. *Annals of Botany*, *101*(6), 747-757.

Frendo, J.-L., Olivier, D., Cheynet, V., Blond, J.-L., Bouton, O., Vidaud, M. *et al.* (2003). Direct involvement of HERV-W *env* glycoprotein in human trophoblast cell fusion and differentiation. *Molecular and Cellular Biology*, *23*(10), 3566-3574.

Freudenstein, J. V. & Chase, M. W. (2001). Analysis of mitochondrial *nad1*b-c intron sequences in Orchidaceae: utility and coding of length-change characters. *Systematic Botany*, *26*(3), 643-657.

Friesen, N., Brandes, A. & Heslop-Harrison, J. S. (Pat). (2001). Diversity, origin, and distribution of retrotransposons (*gypsy* and *copia*) in conifers. *Molecular Biology and Evolution*, *18*(7), 1176-1188.

Frohlich, M. W. & Chase, M. W. (2007). After a dozen years of progress the origin of angiosperms is still a great mystery. *Nature*, *450*(7173), 1184-1189.

Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, *3*(9), 722-732.

Fryirs, C., Eisenhaur, B. & Duckworth, S. (2008). Luteins in lupins — An eye for health. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 488-495). Fremantle, Australia.

Fychan, R., Marley, C., Lewis, G., Davies, R., Theobald, V., Jones, R. *et al.* (2008). Effects of feeding concentrate diets containing narrow-leafed lupin, yellow lupin or soya when compared with a control diet on the productivity of finishing lambs. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 127-130). Fremantle, Australia.

Gabriel, A., Willems, M., Mules, E. H. & Boeke, J. D. (1996). Replication infidelity during a single cycle of Ty*1* retrotransposition. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(15), 7767-7771.

Gadberry, M. D., Malcomber, S. T., Doust, A. N. & Kellogg, E. A. (2005). Primaclade–a

flexible tool to find conserved PCR primers across multiple species. *Bioinformatics*, *21*(7), 1263-1264.

Galbraith, D. W., Harkins, K. R., Maddox, J. M., Ayres, N. M., Sharma, D. P. & Firoozabady, E. (1983). Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science*, *220*(4601), 1049-1051.

Galtier, N. & Jean-Marie, A. (2004). Markov-modulated markov chains and the covarion process of molecular evolution. *Journal of Computational Biology*, *11*(4), 727-733.

Garzón-de la Mora, P., Moreno-Sandoval, L. E., Villáfan-Bernal, J. R., Avalos-Alcantara, G., Gurrola-Díaz, C. M. & García-López, P. M. (2008). *Lupinus albus* seed globulins induce hypoglycaemia and hypotriglyceridemia in Wistar rats. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 469-472). Fremantle, Australia.

Garzón-de la Mora, P., Villáfan-Bernal, J. R., Moreno-Sandoval, L. E., Avalos-Alcantara, G., Gurrola-Díaz, C. M. & García-López, P. M. (2008). *Lupinus exaltatus* alkaloids induces hypoglycemia and hypotriglyceridemia in normal and diabetic Wistar rats. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 529-532). Fremantle, Australia.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685-695.

Gaut, B. S. & Ross-Ibarra, J. (2008). Selection on major components of angiosperm genomes. *Science*, *320*(5875), 484-486.

Gaut, B. S., Wright, S. I., Rizzon, C., Dvorak, J. & Anderson, L. K. (2007). Recombination: an underappreciated factor in the evolution of plant genomes. *Nature Reviews Genetics*, *8*(1), 77-84.

Gent, D. C. van, Mizuuchi, K. & Gellert, M. (1996). Similarities between initiation of V(D)J recombination and retroviral integration. *Science*, *271*(5255), 1592-1594.

Gentry, A. H. (1988). Changes in plant community diversity and floristic composition on environmental and geographical gradients. *Annals of the Missouri Botanical Garden*, *75*(1), 1-34.

Gerstein, A. C., Chun, H.-J. E., Grant, A. & Otto, S. P. (2006). Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. *PLoS Genetics*, *2*(9), e145.

Geurts, R. & Bisseling, T. (2002). *Rhizobium* nod factor perception and signalling. *Plant Cell*, *14*, S239-S249.

Gherbi, H., Markmann, K., Svistoonoff, S., Estevan, J., Autran, D., Giczey, G. *et al.* (2008). SymRK defines a common genetic basis for plant root endosymbioses with arbuscular mycorrhiza fungi, rhizobia, and *Frankia* bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(12), 4928-4932.

Gill, N. T. & Vear, K. C. (1980). *Agricultural Botany* (3^{e} éd.). London: Gerald Duckworth & Co.

Gladstones, J. S. (1974). *Lupins of the mediterranean region and Africa* (Technical bulletin n^{o} 26). Department of Agriculture of Western Australia. (48 p.)

Gladstones, J. S. (1984). Present situation and potential of mediterranean/african *Lupi-*

nus for crop production. In *Proceedings of the Third International Lupin Conference* (p. 67-85). La Rochelle, France.

Gladstones, J. S. (1998). Distribution, origin, taxonomy, history and importance. In J. S. Gladstones, C. Atkins & J. Hamblin (Eds.), *Lupins as crop plants: biology, production and utilization* (p. 1-37). Wallingford, UK: CAB International.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, *36*(2), 182-198.

González, J., Macpherson, J. M. & Petrov, D. A. (2009). A recent adaptive transposable element insertion near highly conserved developmental loci in *Drosophila melanogaster*. *Molecular Biology and Evolution*, *26*(9), 1949-1061.

Gotea, V. & Makałowski, W. (2006). Do transposable elements really contribute to proteomes? *Trends in Genetics*, *22*(5), 260-267.

Gradstein, F. M. & Ogg, J. G. (2004). Geologic time scale 2004 – why, how, and where next! *Lethaia*, *37*(2), 175-181.

Graham, P. H. & Vance, C. P. (2003). Legumes: importance and constraints to greater use. *Plant Physiology*, *131*(3), 872-877.

Grandbastien, M.-A. (1998). Activation of plant retrotransposons under stress conditions. *Trends in Plant Science*, *5*(3), 181-187.

Grandbastien, M.-A., Audéon, C., Bonnivard, E., Casacuberta, J. M., Chalhoub, A.-P. P., Boulos Costa, Le, Q. H. *et al.* (2005). Stress activation and genomic impact of *Tnt1* retrotransposons in Solanaceae. *Cytogenetic and Genome Research*, *110*(1-4), 229-241.

Grandbastien, M.-A., Lucas, H., Morel, J.-B., Mhiri, C., Vernhettes, S. & Casacuberta, J. M. (1997). The expression of the tobacco *Tnt1* retrotransposon is linked to plant defense responses. *Genetica*, *100*(1-3), 241-252.

Graur, D., Shuali, Y. & Li, W.-H. (1989). Deletions in processed pseudogenes accumulate faster in rodents than in humans. *Journal of Molecular Evolution*, *28*(4), 279-285.

Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews of the Cambridge Philosophical Society*, *76*(1), 65-101.

Gregory, T. R. (2005a). *Animal genome size database.* (`http://www.genomesize.com`)

Gregory, T. R. (2005b). The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Annals of Botany*, *95*(1), 133-146.

Gregory, T. R. (2008). Understanding evolutionary trees. *Evolution: Education and Outreach*, *1*(2), 121-137.

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J. *et al.* (2007). Eukaryotic genome size databases. *Nucleic Acids Research*, *35*(suppl. 1), D332-338.

Greilhuber, J. (1998). Intraspecific variation in genome size: a critical reassessment. *Annals of Botany*, *82*(suppl. 1), 27-35.

Greilhuber, J. (2005). Intraspecific variation in genome size in angiosperms: identifying its existence. *Annals of Botany*, *95*(1), 91-98.

Greilhuber, J. (2008). Cytochemistry and C-values: the less-well-known world of nuclear DNA amounts. *Annals of Botany*, *101*(6), 791-804.

Greilhuber, J., Borsch, T., Müller, K. F., Worberg, A., Porembski, S. & Barthlott, W. (2006). Smallest Angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biology (Stuttgart)*, *8*, 770-777.

Greilhuber, J., Doležel, J., Lysák, M. A. & Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms 'Genome Size' and 'C-Value' to describe nuclear DNA contents. *Annals of Botany*, *95*(1), 255-260.

Gribbon, B. M., Pearce, S. R., Kalendar, R., Schulman, A. H., Paulin, L., Jack, P. *et al.* (1999). Phylogeny and transpositional activity of Ty1-*copia* group retrotransposons in cereal genomes. *Molecular & General Genetics*, *261*(6), 883-891.

Gross, R. (1986). Lupins in the Old and New World — A biological-cultural coevolution. In *Proceedings of the Fourth International Lupin Conference* (p. 244-277). Geraldton, Western Australia.

Grover, C. E., Kim, H., Wing, R. A., Paterson, A. H. & Wendel, J. F. (2004). Incongruent patterns of local and global genome size evolution in cotton. *Genome Research*, *14*(8), 1474-1482.

Guillon, F. & Champ, M. M.-J. (2002). Carbohydrate fractions of legumes: uses in human nutrition and potential for health. *British Journal of Nutrition*, *88*(Suppl. 3), 293-306.

Guindon, S. (2003). *Méthodes et algorithmes pour l'approche statistique en phylogénie*. Thèse de doctorat, université de Montpellier II.

Guindon, S., Delsuc, F., Dufayard, J.-F. & Gascuel, O. (2009). Estimating maximum likelihood phylogenies with phyML. In D. Posada (Ed.), *Bioinformatics for DNA Sequence Analysis* In *Methods in Molecular Biology* (Vol. 537, p. 113-137). Humana Press.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, *52*(5), 696-704.

Guindon, S., Lethiec, F., Duroux, P. & Gascuel, O. (2005). PhyML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research*, *33*(suppl. 2), W557-559.

Gupta, S., Buirchell, B. J. & Cowling, W. A. (1996). Interspecific reproductive barriers and genomic similarity among the rough-seeded *Lupinus* species. *Plant Breeding*, *115*(2), 123-127.

Gurrola-Díaz, C. M., Borelli, M. I., Przybyl, A. K., García-López, J. M., Garzón-de la Mora, P. & García-López, P. M. (2008). Insulin secretion effect of 2,17-dioxosparteine, 17-thionosparteine, multiflorine and 17-hydroxy-lupanine on rat Langerhan's islets. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 484-487). Fremantle, Australia.

Haffer, J. (2008). Hypotheses to explain the origin of species in Amazonia. *Brazilian Journal of Biology*, *68*(4), 917-947.

Hall, T. A. (1999-2005). *BioEdit: a user-friendly biological sequence alignment editor and analysis program.* Ibis Therapeutics. Carlsbad, California. (version 7.0.5, `http://www.mbio.ncsu.edu/BioEdit/bioedit.html`)

Hampl, V., Hug, L., Leigh, J. W., Dacks, J. B., Lang, B. F., Simpson, A. G. B. *et al.* (2009).

Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings of the National Academy of Sciences of the United States of America*, *106*(10), 3859-3864.

Harrison, C. J. & Langdale, J. A. (2006). A step by step guide to phylogeny reconstruction. *Plant Journal*, *45*(4), 561-572.

Hawkins, J. S., Hu, G., Rapp, R. A., Grafenberg, J. L. & Wendel, J. F. (2008). Phylogenetic determination of the pace of transposable element proliferation in plants: copia and LINE-like elements in *Gossypium*. *Genome*, *51*(1), 11-18.

Hawkins, J. S., Kim, H., Nason, J. D., Wing, R. A. & Wendel, J. F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Research*, *16*(10), 1252-1261.

Hawkins, J. S., Proulx, S. R., Rapp, R. A. & Wendel, J. F. (2009). Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(42), 17811-17816.

Hennig, U., Hackl, W., Priepke, A., Tuchscherer, A., Souffrant, W. B. & Metges, C. C. (2008). In pigs the true ileal digestibility of essential amino acids of lupin and wheat are equal but the induced ileal basal endogenous losses are different. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 100-104). Fremantle, Australia.

Hennig, W. (1950). *Grundzüge einer Theorie der phylogenetischen Systematik*. Deutscher Zentralverlag, Berlin. (Traduit de l'allemand par D. Dwight Davis et Rainer Zangerl, *Phylogenetic systematics*, University of Illinois Press, Urbana, 1966. Réimprimé en 1979)

Hennig, W. (1969). *Die Stammesgeschichte der Insekten*. Kramer, Frankfurt. (Traduit de l'allemand par Adrian C. Pont, *Insect Phylogeny*, John Wiley & Son, New York, 1981.)

Hetherington, A. M. & Woodward, F. I. (2003). The role of stomata in sensing and driving environmental change. *Nature*, *424*(6951), 901-908.

Heyn, C. C. & Herrnstadt, I. (1977). Seed coat structure of Old World *Lupinus* species. *Botaniska Notiser*, *130*(4), 427-435.

Hill, P., Burford, D., Martin, D. M. & Flavell, A. J. (2005). Retrotransposon populations of *Vicia* species with varying genome size. *Molecular Genetics and Genomics*, *273*(5), 371-381.

Hillis, D. M. & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, *42*(2), 182-192.

Hinegardner, R. (1968). Evolution of cellular DNA content in teleost fishes. *American Naturalist*, *102*(928), 517-523.

Hinegardner, R. (1976). Evolution of genome size. In F. J. Ayala (Ed.), *Molecular evolution* (p. 179-199). Sunderland, Massachusetts: Sinauer.

Hirochika, H., Okamoto, H. & Kakutani, T. (2000). Silencing of retrotransposons in *Arabidopsis* and reactivation by the *ddm1* mutation. *Plant Cell*, *12*(3), 357-369.

Hodgson, J. & Lee, Y. P. (2008). Potential for benefit of lupin on obesity and cardiovascular disease risk in humans. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of*

the 12th International Lupin Conference — Lupins for health and wealth (p. 466-468). Fremantle, Australia.

Hollis, G. F., Hieter, P. A., McBride, O. W., Swan, D. & Leder, P. (1982). Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. *Nature*, *296*(5855), 321-325.

Hollister, J. D. & Gaut, B. S. (2007). Population and evolutionary dynamics of helitron transposable elements in *Arabidopsis thaliana*. *Molecular Biology and Evolution*, *24*(11), 2515-2524.

Holsters, M. (2008). SymRK, an enigmatic receptor guarding and guiding microbial endosymbioses with plant roots. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(12), 4537-4538.

Howieson, J. G., Fillery, I. R. P., Legocki, A., Sikorski, M. M., Stępkowski, T., Minchin, F. R. *et al.* (1998). Nodulation, nitrogen fixation and nitrogen balance. In J. S. Gladstones, C. Atkins & J. Hamblin (Eds.), *Lupins as crop plants: biology, production and utilization* (p. 149-180). CAB International, Wallingford, United kingdom.

Howieson, J. G. & O'Hara, G. W. (2008). Nitrogen fixation by lupins in Western Australia: Which microbes are responsible, from where did they originate, and can we intercede? In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 47-50). Fremantle, Australia.

Huang, X. & Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research*, *9*(9), 868-877.

Hudson, M. E., Lisch, D. R. & Quail, P. H. (2003). The *FHY3* and *FAR1* genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. *Plant Journal*, *34*(4), 453-471.

Huelsenbeck, J. P. & Ronquist, F. (2001). MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754-755.

Hughes, C. E. & Eastwood, R. (2006). Island radiation on a continental scale: Exceptional rates of plant diversification after uplift of the Andes. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(27), 10334-10339.

Hurst, L. D. (2002). The K_A/K_S ratio: diagnosing the form of sequence evolution. *Trends in Genetics*, *18*(9), 486-487.

Jackson, M. (Ed.). (1998). Special issue: Genome Size in Plants. *Annals of Botany*, *82*(suppl A).

Jackson, M. (Ed.). (2005). Special issue: Plant Genome Size. *Annals of Botany*, *95*(1).

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A. *et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463-467.

Jaillon, O., Aury, J.-M. & Wincker, P. (2009). "Changing by doubling", the impact of Whole Genome Duplications in the evolution of eukaryotes. *Comptes Rendus Biologies*, *332*(2-3), 241-253.

Jakob, S. S., Meister, A. & Blattner, F. R. (2004). The considerable genome size variation of *Hordeum* species (Poaceae) is linked to phylogeny, life form, ecology, and speciation rates. *Molecular Biology and Evolution*, *21*(5), 860-869.

Jansen, R. K., Wojciechowski, M. F., Sanniyasi, S.-B., Elumalai Lee & Daniell, H. (2008).

Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (leguminosae). *Molecular Phylogenetics and Evolution*, *48*(3), 1204-1217.

Jansson, S. & Douglas, C. J. (2007). *Populus*: a model system for plant biology. *Annual Review of Plant Biology*, *58*(1), 435-458.

Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S. R., McCouch, S. R. *et al.* (2003). An active DNA transposon family in rice. *Nature*, *421*(6919), 163-167.

Johnson, L. J. (2008). Selfish genetic elements favor the evolution of a distinction between soma and germline. *Evolution*, *62*(8), 2122-2124.

Jones, D. A. & Jones, J. D. G. (1997). The role of leucine-rich repeat proteins in plant defences. In J. H. Andrews, I. C. Tommerup & J. A. Callow (Eds.), *Incorporating Advances in Plant Pathology* In *Advances in Botanical Research* (Vol. 24, p. 89-167). Academic Press.

Jordan, I. K. (2006). Evolutionary tinkering with transposable elements. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(21), 7941-7942.

Josefsson, C., Dilkes, B. & Comai, L. (2006). Parent-dependent loss of gene silencing during interspecies hybridization. *Current Biology*, *16*(13), 1322-1328.

Jovtchev, G., Schubert, V., Meister, A., Barow, M. & Schubert, I. (2006). Nuclear DNA content and nuclear and cell volume are positively correlated in angiosperms. *Cytogenetic and Genome Research*, *114*, 77-82.

Kajita, T., Ohashi, H., Tateishi, Y., Bailey, C. D. & Doyle, J. J. (2001). *rbc*L and Legume phylogeny, with particular reference to Phaseoleae, Millettieae, and allies. *Systematic Botany*, *26*(3), 515-536.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. & Schulman, A. H. (2000). Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE*-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6603-6607.

Kamm, A., Doudrick, R. L., Heslop-Harrison, J. S. & Schmidt, T. (1996). The genomic and physical organization of Ty1-*copia*-like sequences as a component of large genomes in *Pinus elliottii* var. *elliottii* and other gymnosperms. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(7), 2708-2713.

Kapitonov, V. V. & Jurka, J. (2005). RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. *PLoS Biology*, *3*(6), e181.

Kapitonov, V. V. & Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, *9*(5), 411-412.

Kasprzak, A., Šafář, J., Janda, J., Doležel, J., Wolko, B. & Naganowska, B. (2006). The bacterial artificial chromosome (BAC) library of the narrow-leafed lupin (*Lupinus angustifolius* L.). *Cellular & Molecular Biology Letters*, *11*(3), 396-407.

Käss, E. & Wink, M. (1997a). Molecular phylogeny and phylogeography of *Lupinus* (Leguminosae) inferred from nucleotide sequences of the *rbc*L gene and ITS 1+2 regions of rDNA. *Plant Systematics and Evolution*, *208*(3-4), 139-167.

Käss, E. & Wink, M. (1997b). Phylogenetic relationships in the Papilionoideae (Family Leguminosae) based on nucleotide sequences of cpDNA (*rbc*L) and ncDNA (ITS 1 and 2). *Molecular Phylogenetics and Evolution*, *8*(1), 65-88.

Katoh, K., Asimenos, G. & Toh, H. (2009). Multiple Alignment of DNA Sequences with MAFFT. In D. Posada (Ed.), *Bioinformatics for DNA Sequence Analysis* In *Methods in Molecular Biology* (Vol. 537, p. 39-64). Humana Press.

Katoh, K., Kuma, K.-i., Toh, H. & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, *33*(2), 511-518.

Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059-3066.

Katoh, K. & Toh, H. (2007). PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. *Bioinformatics*, *23*(3), 372-374.

Katoh, K. & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, *9*(4), 286-298.

Kazazian, H. H., Jr. (2004). Mobile elements: drivers of genome evolution. *Science*, *303*(5664), 1626-1632.

Kazimierski, T. (1988). An attempt to present lupin evolution of the old world. Materials and impressions. In *Proceedings of the Fifth International Lupin Conference* (p. 103-109). Poznań, Pologne.

Keeling, P. J., Burger, G., Durnford, D. G., Lang, B. F., Lee, R. W., Pearlman, R. E. *et al.* (2005). The tree of eukaryotes. *Trends in Ecology & Evolution*, *20*(12), 670-676.

Kellogg, E. A. & Bennetzen, J. L. (2004). The evolution of nuclear genome structure in seed plants. *American Journal of Botany*, *91*(10), 1709-1725.

Kerbellec, G. (2008). *Apprentissage d'automates modélisant des familles de séquences protéiques*. Thèse de doctorat, Institut national de recherche en informatique et automatisme. (N° d'ordre : 3746, 138 p.)

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49-63.

Kistner, C. & Parniske, M. (2002). Evolution of signal transduction in intracellular symbiosis. *Trends in Plant Science*, *7*(11), 511-518.

Kistner, C., Winzer, T., Pitzschke, A., Mulder, L., Sato, S., Kaneko, T. *et al.* (2005). Seven *Lotus japonicus* genes required for transcriptional reprogramming of the root during fungal and bacterial symbiosis. *Plant Cell*, *17*(8), 2217-2229.

Kjøller, A. & Ødum, S. (1971). Evidence for longevity of seeds and microorganisms in permafrost. *Arctic*, *24*(3), 230-233.

Kłos, P., Poręba, E., Springer, E., Lampart-Szczapa, E. & Goździcka-Józefiak, A. (2008). The potential of narrow-leafed lupin proteins for binding specific and unspecific IgEs from human sera. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 526-528). Fremantle, Australia.

Kluge, A. G. (1998). Total evidence or taxonomic congruence: cladistics or consensus classification. *Cladistics*, *14*(2), 151-158.

Knight, Ch. A. & Ackerly, D. D. (2002). Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters*, *5*(1), 66-76.

Knight, Ch. A. & Beaulieu, J. M. (2008). Genome Size Scaling through Phenotype Space. *Annals of Botany*, *101*(6), 759-766.

Knight, Ch. A., Molinari, N. A. & Petrov, D. A. (2005). The large genome constraint hypothesis: evolution, ecology and phenotype. *Annals of Botany*, *95*(1), 177-190.

Knuth, D. E. (1984). *The TEXbook*. Addison-Wesley, Reading, Massachusetts. (Traduit de l'anglais – États-Unis – par Jean-Côme Charpentier, *Le TEXbook*, Vuibert, 2003.)

Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. *Science*, *304*(5673), 982.

Kobe, B. & Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology*, *11*(6), 725-732.

Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, *7*(1), 474.

Kosakovsky Pond, S. L. & Frost, S. D. W. (2005a). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, *21*(10), 2531-2533.

Kosakovsky Pond, S. L. & Frost, S. D. W. (2005b). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, *22*(5), 1208-1222.

Kosakovsky Pond, S. L., Frost, S. D. W. & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, *21*(5), 676-679.

Kovařík, A., Pires, J. C., Leitch, A. R., Lim, K. Y., Sherwood, A., Matyasek, R. *et al.* (2005). Rapid concerted evolution of nuclear ribosomal DNA in two *Tragopogon* allopolyploids of recent and recurrent origin. *Genetics*, *169*(2), 931-944.

Książkiewicz, M., Karłowski, W., Yang, H. & Wolko, B. (2008). Physical and genetic analysis of genome region conferring resistance to fungal pathogens in the narrow-leafed lupin. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 263-266). Fremantle, Australia.

Lamarck, J.-B. P. A. de Monet, chevalier de. (1809). *Philosophie zoologique*. Paris: Muséum d'Histoire Naturelle. (Réédité par *Flammarion*, Paris, 1997)

Lamport, L. (1994). *LATEX: a document preparation system* (2e éd.). Addison-Wesley Professional, Reading, Massachusetts.

Landan, G. & Graur, D. (2007). Heads or tails: a simple reliability check for multiple sequence alignments. *Molecular Biology and Evolution*, *24*(6), 1380-1383.

Landan, G. & Graur, D. (2009). Characterization of pairwise and multiple sequence alignment errors. *Gene*, *441*(1-2), 141-147.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* (2001). International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860-921.

Larget, B. & Simon, D. L. (1999). Markov Chasin Monte Carlo algorithms for the bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, *16*(6), 750-759.

Lassmann, T., Frings, O. & Sonnhammer, E. L. L. (2009). Kalign2: high-performance

multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Research*, *37*(3), 858-865.

Laurie, D. A. & Bennett, M. D. (1985). Nuclear DNA content in the genera *Zea* and *Sorghum*. Intergeneric, interspecific and intraspecific variation. *Heredity*, *55*(3), 307-313.

Lavin, M., Herendeen, P. S. & Wojciechowski, M. F. (2005). Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Systematic Biology*, *54*(4), 575-594.

Le, S. Q., Lartillot, N. & Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3965-3976.

Lecointre, G. & Deleporte, P. (2005). Total evidence requires exclusion of phylogenetically misleading data. *Zoologica Scripta*, *34*(1), 101-117.

Lee, C., Grasso, C. & Sharlow, M. F. (2002). Multiple sequence alignment using partial order graphs. *Bioinformatics*, *18*(3), 452-464.

Lee, J. Y., Ji, Z. & Tian, B. (2008). Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3′-end of genes. *Nucleic Acids Research*, *36*(17), 5581-5590.

Lee, Y. P., Mori, T. A., Puddey, I. B., Sipsas, S., Ackland, T. R., Beilin, L. J. *et al.* (2009). Effects of lupin kernel flour-enriched bread on blood pressure: a controlled intervention study. *American Journal of Clinical Nutrition*, *89*(3), 766-772.

Leitch, A. R. & Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science*, *320*(5875), 481-483.

Leitch, I. J., Beaulieu, J. M., Cheung, K., Hanson, L., Lysák, M. A. & Fay, M. F. (2007). Punctuated genome size evolution in Liliaceae. *Journal of Evolutionary Biology*, *20*(6), 2296-2308.

Leitch, I. J., Chase, M. W. & Bennett, M. D. (1998). Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. *Annals of Botany*, *82*(suppl. 1), 85-94.

Leitch, I. J., Soltis, D. E., Soltis, P. S. & Bennett, M. D. (2005). Evolution of DNA amounts across land plants (Embryophyta). *Annals of Botany*, *95*(1), 207-217.

Lemaitre, B., Ronsseray, S. & Coen, D. (1993). Maternal repression of the *P* element promoter in the germline of *Drosophila melanogaster*: a model for the *P* cytotype. *Genetics*, *135*(1), 149-160.

Lengyel, S., Gove, A. D., Latimer, A. M., Majer, J. D. & Dunn, R. R. (2009). Ants sow the seeds of global diversification in flowering plants. *PLoS ONE*, *4*(5), e5480.

Le Novère, N. (2001). MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics*, *17*(12), 1226-1227.

Lewis, G., Schrirer, B., Mackinder, B. & Lock, M. (Eds.). (2005). *Legumes of the World*. Royal Botanic Gardens, Kew, London, United Kingdom. (592 p.)

Li, Q.-B. & Guy, Ch. L. (1996). Prolonged final extension time increases cloning efficiency of PCR products. *BioTechniques*, *21*(2), 192-196.

Li, S. & Chou, H.-H. (2004). LUCY2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics*, *20*(16), 2865-2866.

Li, W., Zhang, P., Fellers, J. P., Friebe, B. & Gill, B. S. (2004). Sequence composition, organization, and evolution of the core Triticeae genome. *Plant Journal*, *40*(4), 500-511.

Lim, K. Y., Kovařík, A., Matyasek, R., Chase, M. W., Clarkson, J. J., Grandbastien, M.-A. *et al.* (2007). Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. *New Phytologist*, *175*(4), 756-763.

Linder, C. R., Goertzen, L. R., Vanden Heuvel, B., Francisco-Ortega, J. & Jansen, R. K. (2000). The complete External Transcribed Spacer of 18S-26S rDNA: amplification and phylogenetic utility at low taxonomic levels in Asteraceae and closely allied families. *Molecular Phylogenetics and Evolution*, *14*(2), 285-303.

Liu, B. & Wendel, J. F. (2000). Retrotransposon activation followed by rapid repression in introgressed rice plants. *Genome*, *43*(5), 874-880.

Liu, K., Raghavan, S., Nelesen, S., Linder, C. R. & Warnow, T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, *324*(5934), 1561-1564.

Lohe, A. R. & Hartl, D. L. (1996). Autoregulation of *mariner* transposase activity by overproduction and dominant-negative complementation. *Molecular Biology and Evolution*, *13*(4), 549-555.

Lohe, A. R., Sullivan, D. T. & Hartl, D. L. (1996). Subunit interactions in the *mariner* transposase. *Genetics*, *144*(3), 1087-1095.

Lönnig, W.-E. & Saedler, H. (1997). Plant transposons: contributors to evolution? *Gene*, *205*(1–2), 245-253.

Loÿtynoja, A. & Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(30), 10557-10562.

Loÿtynoja, A. & Goldman, N. (2008a). A model of evolution and structure for multiple sequence alignment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3913-3919.

Loÿtynoja, A. & Goldman, N. (2008b). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, *320*(5883), 1632-1635.

Loÿtynoja, A. & Goldman, N. (2009). Uniting alignments and trees. *Science*, *324*(5934), 1528-1529.

Lu, G., Zhang, S. & Fang, X. (2008). An improved string composition method for sequence comparison. *BMC Bioinformatics*, *9*(suppl. 6), S15.

Lunter, G., Miklós, I., Drummond, A. J., Jensen, J. L. & Hein, J. (2005). Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, *6*(1), 83.

Lynch, M. (2007a). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(suppl. 1), 8597-8604.

Lynch, M. (2007b). *The Origins of Genome Architecture*. Sinauer, Sunderland, Massachusetts. (389 pages)

Lynch, M. & Conery, J. S. (2003). The origins of genome complexity. *Science*, *302*(5649), 1401-1404.

Lysák, M. A., Koch, M. A., Beaulieu, J. M., Meister, A. & Leitch, I. J. (2009). The dynamic ups and downs of genome size evolution in *Brassicaceae*. *Molecular Biology and Evolution*, *26*(1), 85-98.

Lysák, M. A., Koch, M. A., Pecinka, A. & Schubert, I. (2005). Chromosome triplication found across the tribe *Brassiceae*. *Genome Research*, *15*(4), 516-525.

Ma, J., Devos, K. M. & Bennetzen, J. L. (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Research*, *14*(5), 860-869.

Macas, J. & Neumann, P. (2007). Ogre elements — A distinct group of plant Ty3/*gypsy*-like retrotransposons. *Gene*, *390*(1-2), 108-116.

Maciel, H. S. & Schifino-Wittmann, M. T. (2002). First chromosome number determinations in south-eastern South American species of *Lupinus* L. (Leguminosae). *Botanical Journal of the Linnean Society*, *139*(4), 395-400.

Magallón, S. & Castillo, A. (2009). Angiosperm diversification through time. *American Journal of Botany*, *96*(1), 349-365.

Magni, C., Sessa, F., Accardo, E., Vanoni, M., Morazzoni, P., Scarafoni, A. *et al.* (2004). Conglutin γ, a lupin seed protein, binds insulin in vitro and reduces plasma glucose levels of hyperglycemic rats. *Journal of nutritional biochemistry*, *15*(11), 646-650.

Mahé, F., Markova, D., Misset, M.-Th. & Aïnouche, A. (en préparation). Isolation, phylogeny and evolution of the *SymRK* gene in the legume genus *Lupinus* L.

Mahé, F., Pascual, H., Coriton, O., Huteau, V., Navarro Perris, A., Misset, M.-Th. *et al.* (accepté). New data and phylogenetic placement of the recently described Old World lupine species, *Lupinus mariae-josephi*. *Genetic Resources and Crop Evolution*.

Marie, D. & Brown, S. C. (1993). A cytometric exercise in plant DNA histograms, with 2C values for 70 species. *Biology of the Cell*, *78*(1-2), 41-51.

Markmann, K., Giczey, G. & Parniske, M. (2008). Functional adaptation of a plant receptor-kinase paved the way for the evolution of intracellular root symbioses with Bacteria. *PLoS Biology*, *6*(3), e68.

Markmann, K. & Parniske, M. (2009). Evolution of root endosymbiosis with bacteria: how novel are nodules? *Trends in Plant Science*, *14*(2), 77-86.

Marley, C., Davies, D., Fisher, B., Fychan, R., Sanderson, R., Jones, R. *et al.* (2008). Effects of incorporating yellow lupins into concentrate diets compared with soya on milk production and milk composition when offered to dairy cows. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 115-117). Fremantle, Australia.

Martin, W., Roettger, M. & Lockhart, P. J. (2007). A reality check for alignments and trees. *Trends in Genetics*, *23*(10), 478-480.

Martins, J. M., Riottot, M., Abreu, M. C. de, Viegas-Crespo, A. M., Lança, M. J., Almeida, J. A. *et al.* (2005). Cholesterol-lowering effects of dietary blue lupin (*Lupinus angustifolius* L.) in intact and ileorectal anastomosed pigs. *Journal of Lipid Research*, *46*(7), 1539-1547.

Matsubara, K., Kodama, H., Kokubun, H., Watanabe, H. & Ando, T. (2005). Two novel transposable elements in a cytochrome P450 gene govern anthocyanin biosynthesis of commercial petunias. *Gene*, *358*, 121-126.

Matthaei, J. H. & Nirenberg, M. W. (1961). Characteristics and stabilization of DNAase-sensitive protein synthesis in *E. coli* extracts. *Proceedings of the National Academy of Sciences of the United States of America*, *47*(10), 1580-1588.

Matzke, A. J. M. & Matzke, M. A. (1998). Position effects and epigenetic silencing of plant transgenes. *Current Opinion in Plant Biology*, *1*(2), 142-148.

Matzke, M. A. & Matzke, A. J. M. (1998). Epigenetic silencing of plant transgenes as a consequence of diverse cellular defence responses. *Cellular and Molecular Life Sciences*, *54*(1), 94-103.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(11), 344-355.

McClintock, B. (1983, dec 8). *The significance of responses of the genome to challenge.* Nobel lecture.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, *226*(4676), 792-801.

McDougall, I., Brown, F. H. & Fleagle, J. G. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, *433*(7027), 733-736.

McKirdy, S., Shea, G., Hardie, D. & Eagling, D. (2008). Why plant biosecurity? In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 412-415). Fremantle, Australia.

Messier, W. & Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature*, *385*(6612), 151-154.

Meyers, L. A. & Levin, D. A. (2006). On the abundance of polyploids in flowering plants. *Evolution*, *60*(6), 1198-1206.

Mhiri, C. & Grandbastien, M.-A. (2004). Éléments transposables et analyse de la biodiversité végétale. In *La génomique en biologie végétale* (p. 377-402). INRA éditions.

Mi, S., Lee, X., Li, X.-P., Veldman, G. M., Finnerty, H., Racie, L. *et al.* (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, *403*(6771), 785-789.

Michaelson, M. J., Price, H. J., Johnston, J. S. & Ellison, J. R. (1991). Variation of nuclear DNA content in *Helianthus annuus* (Asteraceae). *American Journal of Botany*, *78*(9), 1238-1243.

Mihailović, V., Hill, G. D., Ćupina, B. & Vasiljević, S. (2008). White lupin as a forage crop on alkaline soils. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 79-82). Fremantle, Australia.

Mihailović, V., Hill, G. D., Lazarević, B., Eickmeyer, F., Mikić, A., Krstić, Đ. *et al.* (2008). Performance of blue lupin (*Lupinus angustifolius* L.) cultivars on a pseudogley soil in Serbia. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 51-54). Fremantle, Australia.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H. *et al.* (2008, apr).

The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, *452*(7190), 991-996.

Mirsky, A. E. & Ris, H. (1951). The desoxyribonucleic acid content of animal cells and its evolutionary significance. *Journal of General Physiology*, *34*(4), 451-462.

Misra, S., Buratowski, R. M., Ohkawa, T. & Rio, D. C. (1993). Cytotype control of *Drosophila melanogaster P* element transposition: genomic position determines maternal repression. *Genetics*, *135*(3), 785-800.

Misset, M.-Th. & Gouret, J.-P. (1996). Flow cytometric analysis of the different ploidy levels observed in the genus *Ulex* L. Faboideae-Genisteae in Brittany (France). *Botanica Acta*, *109*(1), 72-79.

Monod, J. (1970). *Le Hasard et la nécessité*. Éditions du Seuil.

Monteiro, R. (1987). Seed testa pattern of unifoliolate species of *Lupinus* L. (Leguminosae). *Salusvita*, *6*, 20-31.

Monteiro, R. & Gibbs, P. E. (1986). A Taxonomic revision of the unifoliolate species of *Lupinus* (Leguminosea) in Brazil. *Notes from the Royal Botanic Garden of Edinburgh*, *44*(1), 71-104.

Moore, B. R. & Donoghue, M. J. (2009). A Bayesian approach for evaluating the impact of historical events on rates of diversification. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(11), 4307-4312.

Morange, M. (1994). *Histoire de la biologie moléculaire*. La Découverte, Paris. (Réédité en 2003)

Moreau, C. S., Bell, C. D., Vila, R., Archibald, S. B. & Pierce, N. E. (2006). Phylogeny of the ants: diversification in the age of angiosperms. *Science*, *312*(5770), 101-104.

Moretti, S., Armougom, F., Wallace, I. M., Higgins, D. G., Jongeneel, C. V. & Notredame, C. (2007). The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Research*, *35*(suppl. 2), W645-648.

Morgenstern, B. (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, *15*(3), 211-218.

Morgenstern, B., Frech, K., Dress, A. & Werner, T. (1998). DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, *14*(3), 290-294.

Morris, W. F. & Wood, D. M. (1989). The role of lupine in succession on Mount St. Helens: facilitation or inhibition? *Ecology*, *70*(3), 697-703.

Morrison, D. A. (2006). Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany*, *19*(6), 479-539.

Morrison, D. A. & Ellis, J. T. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Molecular Biology and Evolution*, *14*(4), 428-441.

Mougel, M., Houzet, L. & Darlix, J.-L. (2009). When is it time for reverse transcription to start and go? *Retrovirology*, *6*(1), 24.

Müller, J. & Müller, K. F. (2004). TreeGraph: automated drawing of complex tree figures using an extensible tree description format. *Molecular Ecology Notes*, *4*(4), 786-788. (Manuel et logiciel disponibles à l'adresse : `http://www.nees.uni-bonn.de/downloads/TreeGraph/tgf.zip`)

Müller, K. F. (2005a). The efficiency of different search strategies in estimating parsimony jackknife, bootstrap, and Bremer support. *BMC Evolutionary Biology*, *5*(1), 1-10.

Müller, K. F. (2005b). SeqState - primer design and sequence statistics for phylogenetic DNA data sets. *Applied Bioinformatics*, *4*(1), 65-69. (`http://www.nees.uni-bonn.de/downloads/SeqState`)

Müller, K. F. (2006). Incorporating information from length-mutational events into phylogenetic analysis. *Molecular Phylogenetics and Evolution*, *38*(3), 667-676.

Naganowska, B., Wolko, B., Śliwińska, E. & Kaczmarek, Z. (2003). Nuclear DNA content variation and species relationships in the genus *Lupinus* (Fabaceae). *Annals of Botany*, *92*(3), 349-355.

Nardon, Ch., Weiss, M., Vieira, C. & Biémont, Ch. (2003). Variation of the genome size estimate with environmental conditions in *Drosophila melanogaster*. *Cytometry Part A*, *55A*(1), 43-49.

Navarro Peris, A., Fos Martín, S., Ferrando Pardo, I. & Laguna Lumbreras, E. (2006). Localización del endemismo aparentemente extinto *Lupinus mariae-josephi*. *Flora Montiberica*, *33*, 59-63.

Neafsey, D. E., Blumenstiel, J. P. & Hartl, D. L. (2004). Different regulatory mechanisms underlie similar transposable element profiles in pufferfish and fruitflies. *Molecular Biology and Evolution*, *21*(12), 2310-2318.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443-453.

Nekrutenko, A. & Li, W.-H. (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends in Genetics*, *17*(11), 619-621.

Nekrutenko, A., Makova, K. D. & Li, W.-H. (2002). The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Research*, *12*(1), 198-202.

Nelesen, S., Liu, K., Zhao, D., Linder, C. R. & Warnow, T. (2008). The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. *Pacific Symposium Biocomputing*, *13*, 25-36.

Nelson, M. N., Boersma, J., Chudy, M., Leśniewska, K., Ellwood, S. R., Phan, H. T. T. *et al.* (2008). A dense reference map of the *Lupinus angustifolius* L. genome: A foundation for building lupin genome research. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 255-258). Fremantle, Australia.

Nelson, M. N., Phan, H. T. T., Ellwood, S. R., Moolhuijzen, P. M., Hane, J., Williams, A. *et al.* (2006). The first gene-based map of *Lupinus angustifolius* L. – location of domestication genes and conserved synteny with *Medicago truncatula*. *Theoretical and Applied Genetics*, *113*(2), 225-238.

Neumann, P., Koblížková, A., Navrátilová, A. & Macas, J. (2006). Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. *Genetics*, *173*(2), 1047-1056.

Neumann, P., Požárková, D. & Macas, J. (2003). Highly abundant pea LTR retrotrans-

poson *Ogre* is constitutively transcribed and partially spliced. *Plant Molecular Biology*, *53*(3), 399-410.

Nirenberg, M. W. & Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, *47*(10), 1588-1602.

Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology*, *302*(1), 205-217.

Novák, Á., Miklós, I., Lyngsø, R. & Hein, J. (2008). StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*, *24*(20), 2403-2404.

Novikova, O., Śliwińska, E., Fet, V., Settele, J., Blinov, A. & Woyciechowski, M. (2007). CR1 clade of non-LTR retrotransposons from *Maculinea* butterflies (Lepidoptera: Lycaenidae): evidence for recent horizontal transmission. *BMC Evolutionary Biology*, *7*(1), 93.

Nowacki, E. (1963). Inheritance and biosynthesis of alkaloids in lupin. *Genetica Polonica*, *4*(2), 161-202.

Nylander, J. A. A. (2004). *MrModeltest v2.* Evolutionary Biology Centre, Uppsala University, Sweden. (version 2.2, programme distribué par l'auteur : `http://people.scs.fsu.edu/~nylander/`)

Nylander, J. A. A., Wilgenbusch, J. C., Warren, D. L. & Swofford, D. L. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, *24*(4), 581-583.

Obbard, D. J., Gordon, K. H. J., Buck, A. H. & Jiggins, F. M. (2009). The evolution of RNAi as a defence against viruses and transposable elements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1513), 99-115.

Obermayer, R., Święcicki, W. K. & Greilhuber, J. (1999). Flow cytometric determination of genome size in some Old World *Lupinus* species (Fabaceae). *Plant Biology (Stuttgart)*, *1*(4), 403-407.

Ogasawara, H., Obata, H., Hata, Y., Takahashi, S. & Gomi, K. (2009). Crawler, a novel Tc1/mariner-type transposable element in *Aspergillus oryzae* transposes under stress conditions. *Fungal Genetics and Biology*.

Ogden, T. H. & Rosenberg, M. S. (2007). How should gaps be treated in parsimony? A comparison of approaches using simulation. *Molecular Phylogenetics and Evolution*, *42*(3), 817-826.

Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag, New York.

Ohno, S. (1972). So much "junk" DNA in our genome. In H. H. Smith (Ed.), *Proceedings of the 23rd Brookhaven Symposium on Biology, session "Evolution of Genetic Systems"* (p. 366-370). Gordon & Breach, New York.

Ohno, S. (1973). Evolutional reason for having so much "junk" DNA. In R. A. Pfeiffer (Ed.), *Modern Aspects of Cytogenetics: Constitutive Heterochromatin in Man* (p. 169-173). F.K. Schattauer Verlag, Stuttgart.

Okamoto, H. & Hirochika, H. (2001). Silencing of transposable elements in plants. *Trends in Plant Science*, *6*(11), 527-534.

Organ, C. L. & Shedlock, A. M. (2009). Palaeogenomics of pterosaurs and the evolution of small genome size in flying vertebrates. *Biology Letters*, *5*(1), 47-50.

Organ, C. L., Shedlock, A. M., Meade, A., Pagel, M. & Edwards, S. V. (2007). Origin of avian genome size and structure in non-avian dinosaurs. *Nature*, *446*(7132), 180-184.

Orgel, L. E. & Crick, F. H. C. (1980). Selfish DNA: the ultimate parasite. *Nature*, *284*(5757), 604-607.

Orlando, L. (2005). *L'Anti-Jurassic Park – Faire parler l'ADN fossile*. Belin.

Östergren, G. (1945). The evolutionary consequences of polyploidy. *Botaniska Notiser*, *2*, 157-163.

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell*, *131*(3), 452-462.

Ouyang, S. & Buell, C. R. (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research*, *32*(suppl. 1), D360-363.

Ovcharenko, I., Loots, G. G., Hardison, R. C., Miller, W. & Stubbs, L. (2004). zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Research*, *14*(3), 472-477.

Page, V., Weisskopf, L. & Feller, U. (2006). Heavy metals in white lupin: uptake, root-to-shoot transfer and redistribution within the plant. *New Phytologist*, *171*(2), 329-341.

Pagel, M. & Meade, A. (2008). Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3955-3964.

Parfrey, L. W., Barbero, E., Lasser, E., Dunthorn, M., Bhattacharya, D., Patterson, D. J. *et al.* (2006, 12). Evaluating support for the current classification of eukaryotic diversity. *PLoS Genetics*, *2*(12), e220.

Parkes, M., Barrett, J. C., Prescott, N. J., Tremelling, M., Anderson, C. A., Fisher, S. A. *et al.* (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genetics*, *39*(7), 830-832.

Parkin, I. A. P., Gulden, S. M., Sharpe, A. G., Lukens, L., Trick, M., Osborn, T. C. *et al.* (2005). Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. *Genetics*, *171*(2), 765-781.

Pascual, H. (2004). *Lupinus mariae-josephi* (Fabaceae), nueva y sorprendente especie descubierta en España. *Anales del Jardín Botánico de Madrid*, *61*(1), 69-72.

Pawlowski, K. & Sprent, J. I. (2008). Comparison between actinorhizal and legume symbiosis. In K. Pawlowski & W. E. Newton (Eds.), *Nitrogen-fixing Actinorhizal Symbioses* In *Nitrogen Fixation: Origins, Applications, and Research Progress* (Vol. 6, p. 261-288). Springer Netherlands.

Pazy, B., Heyn, C. C., Herrnstadt, I. & Plitmann, U. (1977). Studies in populations of the Old World Lupinus species. I. Chromosomes of the East-Mediterranean lupines. *Israel Journal of Botany*, *26*, 115-127.

Peer, Y. Van de, Maere, S. & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, *10*(10), 725-732.

Peeters, K. A. B. M., Nordlee, J. A., Penninks, A. H., Chen, L., Goodman, R. E., Bruijnzeel-Koomen, C. A. F. M. *et al.* (2007). Lupine allergy: not simply cross-reactivity with peanut or soy. *Journal of Allergy and Clinical Immunology*, *120*(3), 647-653.

Peña, T. Coba de la & Sánchez-Moreiras, A. M. (2001). Flow cytometry: cell cycle. In M. J. Reigosa Roger (Ed.), *Handbook of Plant Ecophysiology Techniques* (p. 65-80). Springer Netherlands.

Peñalosa, J. M., Carpena, R. O., Vázquez, S., Agha, R., Granado, A., Sarro, M. J. *et al.* (2007). Chelate-assisted phytoextraction of heavy metals in a soil contaminated with a pyritic sludge. *Science of The Total Environment*, *378*(1-2), 199-204.

Peterson, B. K., Hare, E. E., Iyer, V. N., Storage, S., Conner, L., Papaj, D. R. *et al.* (2009). Big genomes facilitate the comparative identification of regulatory elements. *PLoS ONE*, *4*(3), e4688.

Petit, N. & Barbadilla, A. (2009). Selection efficiency and effective population size in *Drosophila* species. *Journal of Evolutionary Biology*, *22*(3), 516-526.

Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *Trends in Genetics*, *17*(1), 23-28.

Petrov, D. A., Sangster, T. A., Johnston, J. S., Hartl, D. L. & Shaw, K. L. (2000). Evidence for DNA loss as a determinant of genome size. *Science*, *287*(5455), 1060-1062.

Petrov, D. A. & Wendel, J. F. (2006). Evolution of eukaryotic genome structure. In C. W. Fox & J. B. Wolf (Eds.), *Evolutionary genetics: Concepts and case studies* (chap. 10). Oxford University Press.

Petterson, D. S. (1998). Composition and food uses of lupins. In J. S. Gladstones, C. Atkins & J. Hamblin (Eds.), *Lupins as crop plants: biology, production and utilization* (p. 354-383). Wallingford, UK: CAB International.

Petterson, D. S. & Harris, D. J. (1995). Cadmium and lead content of lupin seed grown in Western Australia. *Australian Journal of Experimental Agriculture*, *35*(3), 403-407.

Peyret, N., Seneviratne, P. A., Allawi, H. T. & SantaLucia, J., Jr. (1999). Nearest-Neighbor Thermodynamics and NMR of DNA sequences with internal A·A, C·C, G·G, and T·T mismatches. *Biochemistry*, *38*(12), 3468-3477.

Pfeil, B. E., Schlueter, J. A., Shoemaker, R. C. & Doyle, J. J. (2005). Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in Legumes using 39 gene families. *Systematic Biology*, *54*(3), 441-454.

Phan, H. T., Ellwood, S. R., Adhikari, K., Nelson, M. N. & Oliver, R. P. (2007). The first genetic and comparative map of white lupin (*Lupinus albus* L.): identification of QTLs for anthracnose resistance and flowering time, and a locus for alkaloid content. *DNA Research*, *14*(2), 59-70.

Phillips, L. L. (1957). Chromosome numbers in *Lupinus*. *Madroño*, *14*, 30-36.

Piégu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H. *et al.* (2006). Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research*, *16*(10), 1262-1269.

Pilvi, T. K., Jauhiainen, T., Cheng, Z. J., Mervaala, E. M., Vapaatalo, H. & Korpela, R. (2006). Lupin protein attenuates the development of hypertension and norma-

lises the vascular function of NaCl-loaded Goto-Kakizaki rats. *Journal of Physiology and Pharmacology*, *57*(2), 167-176.

Piotrowicz-Cieślak, A. I., Adomas, B., Michalczyk, D. J. & Górska, K. (2008). Ultrastructural analysis of seed testa in developing *Lupinus pilosus* seeds. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 192-194). Fremantle, Australia.

Planchuelo, A. M. (1984). Taxonomic studies of *Lupinus* in South America. In *Proceedings of the Third International Lupin Conference* (p. 39-54). La Rochelle, France.

Planchuelo, A. M. (1994). Wild lupins distribution and its implication as germplasm resources. In J. M. Neves Martins & M. L. Beirao da Costa (Eds.), *Proceedings of the 7th International Lupin Conference — Genetic resources & plant breeding* (p. 65-69). Evora, Portugal.

Planchuelo, A. M. & Dunn, D. B. (1984). The simple leaved lupines and their relatives in Argentina. *Annals of the Missouri Botanical Garden*, *71*(1), 92-103.

Plitmann, U. & Heyn, C. C. (1984). Old World *Lupinus*: Taxonomy, evolutionary relationships and links with New World species. In *Proceedings of the Third International Lupin Conference* (p. 55-66). La Rochelle, France.

Plitmann, U. & Pazy, B. (1984). Cytogeographical distribution of the Old World *Lupinus*. *Webbia*, *38*, 531-539.

Pol, D. & Siddall, M. E. (2001). Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. *Cladistics*, *17*(3), 266-281.

Polhill, R. M. (1976). *Genisteae (Adans.) Benth. and related tribes (Leguminosae)*. Academic Press, London. (368 p.)

Poon, A. F. Y., Frost, S. D. W. & Kosakovsky Pond, S. L. (2009). Detecting signatures of selection from DNA sequences using Datamonkey. In D. Posada (Ed.), *Bioinformatics for DNA Sequence Analysis* In *Methods in Molecular Biology* (Vol. 537, p. 163-183). Humana Press.

Porsild, A. E., Harington, C. R. & Mulligan, G. A. (1967). *Lupinus arcticus* Wats. grown from seeds of Pleistocene age. *Science*, *158*(3797), 113-114.

Posada, D. (2006). ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Research*, *34*(suppl. 2), W700-703.

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, *25*(7), 1253-1256.

Posada, D. (2009). Selection of models of DNA evolution with jModelTest. In D. Posada (Ed.), *Bioinformatics for DNA Sequence Analysis* In *Methods in Molecular Biology* (Vol. 537, p. 93-112). Humana Press.

Posada, D. & Buckley, Th. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, *53*(5), 793-808.

Posada, D. & Crandall, K. A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics*, *14*(9), 817-818.

Pourtau, N., Lauga, B., Audéon, C., Grandbastien, M.-A., Goulas, Ph. & Salvado, J.-C. (2003). The promoter of the *Tnt1A* retrotransposon is activated by ozone

air pollution in tomato, but not in its natural host tobacco. *Plant Science*, *165*(5), 983-992.

Price, H. J., Sparrow, A. H. & Nauman, A. F. (1973). Correlations between nuclear volume, cell volume and DNA content in meristematic cells of herbaceous angiosperms. *Experientia*, *29*(8), 1028-1029.

Pryer, K. M., Schneider, H., Smith, A. R., Cranfill, R., Wolf, P. G., Hunt, J. S. *et al.* (2001). Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature*, *409*(6820), 618-622.

Przyborowski, J. A. & Weeden, N. F. (2001). RAPD-based assessment of genetic similarity and distance between *Lupinus* species in section *Albus*. *Journal of Applied Genetics*, *42*(4), 425-433.

Przybylska, J. & Zimniak-Przybylska, Z. (1995). Electrophoretic patterns of seed globulins in the Old-World *Lupinus* species. *Genetic Resources and Crop Evolution*, *42*(1), 69-75.

Quaresma, R. R., Viseu, R., Martins, L. M., Tomaz, E. & Inacio, F. (2007). Allergic primary sensitization to lupine seed. *Allergy*, *62*(12), 1473-1474.

Quattrocchio, F., Wing, J., Woude, K. van der, Souer, E., Vetten, N. de, Mol, J. *et al.* (1999). Molecular analysis of the *anthocyanin2* gene of petunia and its role in the evolution of flower color. *Plant Cell*, *11*(8), 1433-1444.

Radutoiu, S., Madsen, L. H., Madsen, E. B., Felle, H. H., Umehara, Y., Grønlund, M. *et al.* (2003). Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature*, *425*(6958), 585-592.

Rakocevic, A., Mondy, S., Tirichine, L., Cosson, V., Brocard, L., Iantcheva, A. *et al.* (2009). MERE1, a low copy number copia-type retroelement in *Medicago truncatula* active during tissue culture. *Plant Physiology*, *151*(3), 1250-1263.

Ramakrishna, W., Dubcovsky, J., Park, Y.-J., Busso, C., Emberton, J., SanMiguel, P. *et al.* (2002). Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics*, *162*(3), 1389-1400.

Ramallo, E., Kalendar, R., Schulman, A. H. & Martínez-Izquierdo, J. A. (2008). *Reme1*, a *copia* retrotransposon in melon, is transcriptionally induced by UV light. *Plant Molecular Biology*, *66*(1-2), 137-150.

Raphael, B., Zhi, D., Tang, H. & Pevzner, P. (2004). A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Research*, *14*(11), 2336-2346.

Rayburn, A. L. & Auger, J. A. (1990). Genome size variation in *Zea mays* ssp. *mays* adapted to different altitudes. *Theoretical and Applied Genetics*, *79*(4), 470-474.

Rayburn, A. L., Price, H. J., Smith, J. D. & Gold, J. R. (1985). C-band heterochromatin and DNA content in *Zea mays*. *American Journal of Botany*, *72*(10), 1610-1617.

R Development Core Team. (2008). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (`http://www.R-project.org`)

Ree, R. H., Citerne, H. L., Lavin, M. & Cronk, Q. C. B. (2004). Heterogeneous selection on LEGCYC paralogs in relation to flower morphology and the phylogeny of

Lupinus (Leguminosae). *Molecular Biology and Evolution*, *21*(2), 321-331.

Reeves, J. H. (1992). Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *Journal of Molecular Evolution*, *35*(1), 17-31.

Reilly, J. G., Ogden, R. & Rossi, J. J. (1982). Isolation of a mouse pseudo tRNA gene encoding CCA — A possible example of reverse flow of genetic information. *Nature*, *300*(5889), 287-289.

Reis, A. M., Fernandes, N. P., Marques, S. L., Paes, M. J., Sousa, S., Carvalho, F. *et al.* (2007). Lupine sensitisation in a population of 1,160 subjects. *Allergologia et Immunopathologia*, *35*(4), 162-163.

Remy, W., Taylor, T. N., Hass, H. & Kerp, H. (1994). Four hundred-million-year-old vesicular arbuscular mycorrhizae. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(25), 11841-11843.

Resta, D., Boschin, G., D'Agostina, A. & Arnoldi, A. (2008). Quantification of quinolizidine alkaloids in lupin seeds, lupin-based ingredients and foods. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 533-535). Fremantle, Australia.

Richardson, A. O. & Palmer, J. D. (2007). Horizontal gene transfer in plants. *Journal of Experimental Botany*, *58*(1), 1-9.

Ridoux, O. & Lesventes, G. (2008). *Calculateurs, calculs, calculabilité*. Dunod. (204 p.)

Riely, B. K., Mun, J.-H. & Ané, J.-M. (2006). Unravelling the molecular basis for symbiotic signal transduction in legumes. *Molecular Plant Pathology*, *7*(3), 197-207.

Ripplinger, J. & Sullivan, J. (2008). Does choice in model selection affect maximum likelihood analysis? *Systematic Biology*, *57*(1), 76-85.

Roch, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *3*(1), 92-94.

Rodrigue, N., Kleinman, C. L., Philippe, H. & Lartillot, N. (2009). Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Molecular Biology and Evolution*, *26*(7), 1663-1676.

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572-1574.

Roulin, A., Piégu, B., Fortuné, Ph. M., Sabot, F., D'Hont, A., Manicacci, D. *et al.* (2009). Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon *Route66* in Poaceae. *BMC Evolutionary Biology*, *9*(1), 58.

Roulin, A., Piégu, B., Wing, R. A. & Panaud, O. (2008). Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon *RIRE1* within the genus *Oryza*. *Plant Journal*, *53*(6), 950-959.

Rozen, S. & Skaletsky, H. J. (2000). Primer3 on the WWW for general users and for biologist programmers. In S. Krawetz & S. Misener (Eds.), *Bioinformatics Methods and Protocols* (p. 365-386). Humana Press, Totowa, New Jersey.

Ruas, C. F., Weiss-Schneeweiss, H., Stuessy, T. F., Samuel, M. R., Pedrosa-Harand, A., Tremetsberger, K. *et al.* (2008). Characterization, genomic organization and chro-

mosomal distribution of Ty1-*copia* retrotransposons in species of *Hypochaeris* (Asteraceae). *Gene*, *412*(1-2), 39-49.

Saarela, J. M., Rai, H. S., Doyle, J. A., Endress, P. K., Mathews, S., Marchant, A. D. *et al.* (2007). Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. *Nature*, *446*(7133), 312-315.

Saito, K., Yoshikawa, M., Yano, K., Miwa, H., Uchida, H., Asamizu, E. *et al.* (2007). NUCLEOPORIN85 is required for calcium spiking, fungal and bacterial symbioses, and seed production in *Lotus japonicus*. *Plant Cell*, *19*(2), 610-624.

Salamov, A. A. & Solovyev, V. V. (2000). Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research*, *10*(4), 516-522.

Salmanowicz, B. P. & Przybylska, J. (1994). Electrophoretic patterns of seed albumins in the Old-World *Lupinus* species (Fabaceae): Variation in the 2S albumin class. *Plant Systematics and Evolution*, *192*(1), 67-78.

Salvo-Garrido, H., Travella, S., Schwarzacher, T., Harwood, W. A. & Snape, J. W. (2001). An efficient method for the physical mapping of transgenes in barley using in situ hybridization. *Genome*, *44*(1), 104-110.

Sambrook, J. & Russell, D. W. (2001). Plasmids and their usefulness in molecular cloning. In N. Irwin & K. A. Janssen (Eds.), *Molecular Cloning — A laboratory manual* (3^{e} éd., Vol. 1, chap. 1). Cold Spring Harbor Laboratory Press, New York.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463-5467.

SanMiguel, P. & Bennetzen, J. L. (1998). Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Annals of Botany*, *82*(suppl A), 37-44.

Santini, F., Harmon, L. J., Carnevale, G. & Alfaro, M. E. (2009). Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evolutionary Biology*, *9*(1), 194.

Sasaki, T., Nishihara, H., Hirakawa, M., Fujimura, K., Tanaka, M., Kokubo, N. *et al.* (2008). Possible involvement of SINEs in mammalian-specific brain formation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(11), 4220-4225.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M. *et al.* (2008). Genome Structure of the Legume, *Lotus japonicus*. *DNA Research*, *15*(4), 227-239.

Schmitt, Th. (2007). Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*, *4*(1), 11.

Schmollinger, M., Nieselt, K., Kaufmann, M. & Morgenstern, B. (2004). DIALIGN P: fast pair-wise and multiple sequence alignment using parallel processors. *BMC Bioinformatics*, *5*(1), 128.

Schmuths, H., Meister, A., Horres, R. & Bachmann, K. (2004). Genome size variation among accessions of *Arabidopsis thaliana*. *Annals of Botany*, *93*(3), 317-321.

Schulman, A. H. & Kalendar, R. (2005). A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. *Cytogenetic and Genome Research*, *110*(1-4), 598-6058.

Schultz, J. & Wolf, M. (2009). ITS2 sequence-structure analysis in phylogenetics: A how-to manual for molecular systematics. *Molecular Phylogenetics and Evolution*, *52*(2), 520-523.

Schüßler, A., Schwarzott, D. & Walker, C. (2001). A new fungal phylum, the Glomeromycota: phylogeny and evolution. *Mycological Research*, *105*(12), 1413-1421.

Schwartz, D. (1963). *Méthodes statistiques à l'usage des médecins et des biologistes* (4e éd.). Flammarion/Médecine-Sciences, Paris. (imprimé en 1995, troisième tirage de la quatrième édition, 314 p.)

Schwarzacher, T. & Heslop-Harrison, J. S. (Pat). (2000). *Practical* in situ *hybridization* (1re éd.). BIOS Scientific Publishers, Oxford, United Kingdom. (605 p.)

Seberg, O. & Petersen, G. (2009). A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews Genetics*, *10*(4), 276.

Seelanan, T., Schnabel, A. & Wendel, J. F. (1997). Congruence and consensus in the cotton tribe (Malvaceae). *Systematic Botany*, *22*(2), 259-290.

Serrano, E., Storebakken, T., Penn, M., Landsverk, T., Hansen, J. Ø. & Mydland, L. T. (2008). Responses in rainbow trout (Oncorhynchus mykiss) to increasing dietary dose of lupinine alkaloid. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 94-99). Fremantle, Australia.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M. & Levy, A. A. (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell*, *13*(8), 1749-1759.

Shapiro, B., Rambaut, A. & Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution*, *23*(1), 7-9.

Sharp, P. A., Sugden, B. & Sambrook, J. (1973). Detection of two restriction endonuclease activities in *Haemophilus parainfluenzae* using analytical agarose-ethidium bromide electrophoresis. *Biochemistry*, *12*(16), 3055-3063.

Shea, G., Thomas, G., Buirchell, B. J., Salam, M., McKirdy, S. & Sweetingham, M. (2008). Case study: industry response to the lupin anthracnose incursion in Western Australia. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 425-431). Fremantle, Australia.

Shi, B.-J., Gustafson, J. P. & Langridge, P. (2009). A simple TAE-based method to generate large insert BAC libraries from plant species. In D. J. Somers, P. Langridge & J. P. Gustafson (Eds.), *Plant Genomics* In *Methods in Molecular Biology* (Vol. 513, p. 1-24). Humana Press.

Shirasu, K., Schulman, A. H., Lahaye, T. & Schulze-Lefert, P. (2000). A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Research*, *10*(7), 908-915.

Shoemaker, R. C., Schlueter, J. & Doyle, J. J. (2006). Paleopolyploidy and gene duplication in soybean and other legumes. *Current Opinion in Plant Biology*, *9*(2),

104-109.

Siddall, M. E. & Whiting, M. F. (1999). Long-branch abstractions. *Cladistics*, *15*(1), 9-24.

Siefert, J. L. (2009). Defining the mobilome. In M. B. Gogarten, J. P. Gogarten & L. C. Olendzenski (Eds.), *Horizontal Gene Transfer: Genomes in Flux* In *Methods in Molecular Biology* (Vol. 532, p. 13-27). Humana Press.

Simmons, M. P., Müller, K. F. & Norton, A. P. (2007). The relative performance of indel-coding methods in simulations. *Molecular Phylogenetics and Evolution*, *44*(2), 724-740.

Simmons, M. P. & Ochoterena, H. (2000). Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology*, *49*(2), 369-381.

Simmons, M. P., Pickett, K. M. & Miya, M. (2004). How meaningful are bayesian support values? *Molecular Biology and Evolution*, *21*(1), 188-199.

Sipsas, S. (2008). Lupin products — Concepts and reality. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 506-513). Fremantle, Australia.

Sirtori, C. R., Lovati, M. R., Manzoni, C., Castiglioni, S., Duranti, M., Magni, C. *et al.* (2004). Proteins of white lupin seed, a naturally isoflavone-poor legume, reduce cholesterolemia in rats and increase LDL receptor activity in HepG2 cells. *Journal of Nutrition*, *134*(1), 18-23.

Sirtori, E., O'Kane, F., Brambilla, F. & Arnoldi, A. (2008). *L. angustifolius* vs *L. albus*: a combined chromatographic and electrophoretic analysis to highlight the differences in protein profile. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 212-215). Fremantle, Australia.

Small, R. L., Cronn, R. C. & Wendel, J. F. (2004). Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany*, *17*(2), 145-170.

Šmarda, P., Bureš, P., Horová, L. & Rotreklová, O. (2008). Intrapopulation genome size dynamics in *Festuca pallens*. *Annals of Botany*, *102*(4), 599-607.

Smith, Ch. P. (1944). Lupinus L. In L. Abrahams (Ed.), *Illustrated Flora of the Pacific States* (Vol. 2, p. 483-519). Standford University Press, California.

Smith, J. D. & Gregory, T. R. (2009). The genome sizes of megabats (Chiroptera: Pteropodidae) are remarkably constrained. *Biology Letters*, *5*(3), 347-351.

Smith, P. M. C., Goggin, D. E., Mir, G. A., Cameron, E., Colinet, H., Stuckey, M. *et al.* (2008). Characterisation of allergenic proteins in lupin seeds and the relationship between peanut and lupin allergens. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 459-462). Fremantle, Australia.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195-197.

Soltis, D. E., Mavrodiev, E. V., Doyle, J. J., Rauscher, J. & Soltis, P. S. (2008). ITS and ETS sequence data and phylogeny reconstruction in allopolyploids and hybrids. *Systematic Botany*, *33*(1), 7-20.

Soltis, D. E. & Soltis, P. S. (1998). Choosing an approach and an appropriate gene for phylogenetic analysis. In D. E. Soltis, P. S. Soltis & J. J. Doyle (Eds.), *Molecular*

systematics of plants II: DNA sequencing (p. 1-42). Kluwer Academic Publishers, Boston, Massachusetts.

Soltis, D. E., Soltis, P. S., Bennett, M. D. & Leitch, I. J. (2003). Evolution of genome size in the angiosperms. *American Journal of Botany*, *90*(11), 1596-1603.

Soltis, D. E., Soltis, P. S., Endress, P. K. & Chase, M. W. (2005). *Angiosperm phylogeny and evolution*. Sinauer, Sunderland, Massachusetts.

Soltis, D. E., Soltis, P. S., Morgan, D. R., Swensen, S. M., Mullin, B. C., Dowd, J. M. *et al.* (1995). Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in Angiosperms. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(7), 2647-2651.

Sorek, R., Ast, G. & Graur, D. (2002). *Alu*-containing exons are alternatively spliced. *Genome Research*, *12*(7), 1060-1067.

Stamatakis, A. & Ott, M. (2008). Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3977-3984.

States, D. J. & Boguski, M. S. (1991). Similarity and homology. In M. R. Gribskov & J. Devereux (Eds.), *Sequence analysis primer* (p. 89-157). Stockton Press, New York, USA.

Steel, M. (2005). Should phylogenetic models be trying to 'fit an elephant'? *Trends in Genetics*, *21*(6), 307-309.

Steinbauerová, V., Neumann, P. & Macas, J. (2008). Experimental evidence for splicing of intron-containing transcripts of plant LTR retrotransposon *Ogre*. *Molecular Genetics and Genomics*, *280*(5), 427-436.

Stępkowski, T., Hughes, C. E., Law, I. J., Markiewicz, Ł., Gurda, D., Chlebicka, A. *et al.* (2007). Diversification of lupin *Bradyrhizobium* strains: evidence from nodulation gene trees. *Applied and Environmental Microbiology*, *73*(10), 3254-3264.

Stevens, P. F. (2001). *Angiosperm Phylogeny Website.* (version 9, juin 2008, dernière mise à jour le 19/02/2009 — `http://www.mobot.org/MOBOT/research/APweb/`)

Stracke, S., Kistner, C., Yoshida, S., Mulder, L., Sato, S., Kaneko, T. *et al.* (2002). A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature*, *417*(6892), 959-962.

Suárez-Santiago, V. N., Salinas, M. J., Garcia-Jacas, N., Soltis, P. S., Soltis, D. E. & Blanca, G. (2007). Reticulate evolution in the *Acrolophus* subgroup (*Centaurea* L., Compositae) from the western Mediterranean: Origin and diversification of section *Willkommia* Blanca. *Molecular Phylogenetics and Evolution*, *43*(1), 156-172.

Subramanian, A. R., Weyer-Menkhoff, J., Kaufmann, M. & Morgenstern, B. (2005). DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, *6*(1), 66.

Suchý, P., Straková, E., Kroupa, L. & Večerek, V. (2008). The fatty acid content of oil from seeds of some lupin varieties. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 188-191). Fremantle, Australia.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics. Theory and Methods*, *7*(1), 13-26.

Suzuki, Y., Glazko, G. V. & Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(25), 16138-16143.

Święcicki, W., Święcicki, W. K. & Wolko, B. (1996). *Lupinus anatolicus* – a new lupin species of the old world. *Genetic Resources and Crop Evolution*, *43*(2), 109-117.

Swift, H. (1950). The constancy of Desoxyribose Nucleic Acid in plant nuclei. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(11), 643-654.

Swigoňová, Z., Bennetzen, J. L. & Messing, J. (2005). Structure and evolution of the *r/b* chromosomal regions in rice, maize and sorghum. *Genetics*, *169*(2), 891-906.

Swofford, D. L. (1989-2003). *Phylogenetic Analysis Using Parsimony (*and other methods).* Sinauer Associates. Sunderland, Massachusetts. (version 4.0b10, `http://paup.csit.fsu.edu/`)

Taberlet, P., Gielly, L., Pautou, G. & Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology*, *17*(5), 1105-1109.

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, *10*(3), 512-526.

Tamura, M., Kajikawa, M. & Okada, N. (2007). Functional splice sites in a zebrafish LINE and their influence on zebrafish gene expression. *Gene*, *390*(1-2), 221-231.

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M. & Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, *320*(5875), 486-488.

Terol, J., Castillo, M. C., Bargues, M., Pérez-Alonso, M. & Frutos, R. de. (2001). Structural and evolutionary analysis of the *copia*-like elements in the *Arabidopsis thaliana* genome. *Molecular Biology and Evolution*, *18*(5), 882-892.

Terol, J., Naranjo, M. A., Ollitrault, P. & Talon, M. (2008). Development of genomic resources for *Citrus clementina*: Characterization of three deep-coverage BAC libraries and analysis of 46,000 BAC end sequences. *BMC Genomics*, *9*(1), 423.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408*(6814), 796-815.

Thomas, Ch. A. (1971). The genetic organization of chromosomes. *Annual Review of Genetics*, *5*, 237-256.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673-4680.

Thompson, K. (1990). Genome size, seed size and germination temperature in herbaceous angiosperms. *Evolutionary Trends in Plants*, *4*(2), 113-116.

Tompa, R., McCallum, C. M., Delrow, J., Henikoff, J. G., Steensel, B. van & Henikoff, S. (2002). Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. *Current Biology*, *12*(1), 65-68.

Tournefort, J. Pitton de. (1700). *Institutiones rei herbariæ* (Vol. 1er). Typographia regia, Parisiis.

Tournefort, J. Pitton de. (1717). *Relation d'un voyage du Levant fait par ordre du Roy* (Vol. 1er). Imprimerie royale, Paris.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. & Kakutani, T. (2009). Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature*, *461*(7262), 423-426.

Turner, B. L. (1957). The chromosomal distributional relationships of *Lupinus texensis* and *L. subcarnosus*. *Madroño*, *14*, 13–16.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U. *et al.* (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, *313*(5793), 1596-1604.

Tutin, T. G. *et al.* (Eds.). (1968-1980). *Flora europaea* (Vol. 1-5). Cambridge University Press.

Urbatsch, L. E., Roberts, R. P. & Karaman, V. (2003). Phylogenetic evaluation of *Xylothamia*, *Gundlachia*, and related genera (Asteraceae, Astereae) based on ETS and ITS nrDNA sequence data. *American Journal of Botany*, *90*(4), 634-649.

Van't Hof, J. & Sparrow, A. H. (1963). A relationship between DNA content, nuclear volume, and minimum mitotic cycle time. *Proceedings of the National Academy of Sciences of the United States of America*, *49*(6), 897-902.

Varadarajan, A., Bradley, R. & Holmes, I. (2008). Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. *Genome Biology*, *9*(10), R147.

Vega, R. de la, Gutierrez, M. P., Sanz, C., Calvo, R., Robredo, L. M., Cuadra, C. de la *et al.* (1996). Bactericid-like effect of *Lupinus* alkaloids. *Industrial Crops and Products*, *5*(2), 141-148.

Veksler-Lublinksy, I., Barash, D., Avisar, C., Troim, E., Chew, P. & Kedem, K. (2008). Fash: A web application for nucleotides sequence search. *Source Code for Biology and Medicine*, *3*(1), 9.

Vendrely, R. & Vendrely, C. (1948). La teneur du noyau cellulaire en acide désoxyribonucléique à travers les organes, les individus et les espèces animales : techniques et premiers résultats. *Experientia*, *4*(44), 434-436.

Vendrely, R. & Vendrely, C. (1950). Sur la teneur absolue en acide désoxyribonucléique du noyau cellulaire chez quelques espèces d'oiseaux et de poissons. *Comptes Rendus de l'Académie des Sciences*, *230*, 788-790.

Veuille, M., Brazier, L. & Meli, F. (2008). Interactions génomiques en régime de sélection et taille de population. In É. Charvolin, F. Fridlansky, F. Marie & A. Saïhi (Eds.), *Actes du Bureau des Ressources Génétiques, 7e colloque national : les ressources génétiques à l'heure des génomes* (p. 423-437). Strasbourg, 13-15 octobre.

Vicient, C. M., Jääskeläinen, M. J., Kalendar, R. & Schulman, A. H. (2001). Active retrotransposons are a common feature of grass genomes. *Plant Physiology*, *125*(3), 1283-1292.

Vicient, C. M., Kalendar, R. & Schulman, A. H. (2001). *Envelope*-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. *Genome Research*, *11*(12), 2041-2049.

Vicient, C. M., Suoniemi, A., Anamthawat-Jónsson, K., Tanskanen, J., Beharav, A.,

Nevo, E. *et al.* (1999). Retrotransposon *BARE*-1 and its role in genome evolution in the genus *Hordeum*. *Plant Cell*, *11*(9), 1769-1784.

Virgile. (I^er^ siècle av. J.-C.). *Georgicon*. (Traduit du latin par Maurice Rat, *Les Bucoliques et les Géorgiques*, Garnier, 1932)

Vitte, C. & Panaud, O. (2003). Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Molecular Biology and Evolution*, *20*(4), 528-540.

Vitte, C., Panaud, O. & Quesneville, H. (2007). LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics*, *8*(1), 218.

Voytas, D. F., Cummings, M. P., Konieczny, A., Ausubel, F. M. & Rodermel, S. R. (1992). *Copia*-like retrotransposons are ubiquitous among plants. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(15), 7124-7128.

Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, *34*(6), 1692-1699.

Wang, H., Moore, M. J., Soltis, P. S., Bell, C. D., Brockington, S. F., Alexandre, R. *et al.* (2009). Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences of the United States of America*, -.

Wang, H.-C., Susko, E. & Roger, A. (2009). PROCOV: maximum likelihood estimation of protein phylogeny under covarion models and site-specific covarion pattern analysis. *BMC Evolutionary Biology*, *9*(1), 225.

Wang, L. & Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, *1*(4), 337-348.

Wassenberg, J. & Hofer, M. (2007). Lupine-induced anaphylaxis in a child without known food allergy. *Annals of allergy, asthma & immunology*, *98*(6), 589-590.

Waugh O'Neill, R. J., O'Neill, M. J. & Marshall Graves, J. A. (1998). Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature*, *393*(6680), 68-72.

Wawrzynski, A., Ashfield, T., Chen, N. W., Mammadov, J., Nguyen, A., Podicheti, R. *et al.* (2008). Replication of nonautonomous retroelements in soybean appears to be both recent and common. *Plant Physiology*, *148*(4), 1760-1771.

Weil, C. F. & Wessler, S. R. (1990). The Effects of plant transposable element insertion on transcription initiation and RNA processing. *Annual Review of Plant Physiology and Plant Molecular Biology*, *41*(1), 527-552.

Weiss-Schneeweiss, H., Greilhuber, J. & Schneeweiss, G. M. (2005). Genome size evolution in holoparasitic *Orobanche* (Orobanchaceae) and related genera. *American Journal of Botany*, *93*(1), 148-156.

Wendel, J. F. (2000). Genomic evolution in polyploids. *Plant Molecular Biology*, *42*(1), 225-249.

Wendel, J. F., Cronn, R. C., Spencer Johnston, J. & James Price, H. (2002). Feast and famine in plant genomes. *Genetica*, *115*(1), 37-47.

Wendel, J. F. & Doyle, J. J. (1998). Phylogenetic incongruence: window into genome history and molecular evolution. In P. S. Soltis, D. E. Soltis & J. J. Doyle (Eds.),

Molecular Systematics of Plants II: DNA sequencing (p. 265-296). Kluwer Academic Publishers, Boston, Massachusetts.

Wessler, S. R., Bureau, T. E. & White, S. E. (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Current Opinion in Genetics & Development*, *5*(6), 814-821.

Whelan, S. (2008). The genetic code can cause systematic bias in simple phylogenetic models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 4003-4011.

Whelan, S., Bakker, P. I. W. de, Quevillon, E., Rodriguez, N. & Goldman, N. (2006). PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Research*, *34*(suppl. 1), D327-D331.

White, Th. J., Bruns, S. B., T. D. Lee & Taylor, J. W. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In M. A. Innis, D. H. Gelfand, J. J. Sminsky & Th. J. White (Eds.), *PCR Protocols: A guide to methods and applications* (p. 315-322). Academic Publishers, San Diego, California.

Whittaker, R. H. (1969). New concepts of Kingdoms of organisms. *Science*, *163*(3863), 150-160.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B. *et al.* (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973-982.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B. *et al.* (2009). Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews Genetics*, *10*(276).

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.-D., Dubcovsky, J. *et al.* (2003). Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. *Plant Cell*, *15*(5), 1186-1197.

Wikström, N., Savolainen, V. & Chase, M. W. (2001). Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society B: Biological Sciences*, *268*(1482), 2211-2220.

Wilf, P., Labandeira, C. C., Kress, W. J., Staines, C. L., Windsor, D. M., Allen, A. L. *et al.* (2000). Timing the radiations of leaf beetles: hispines on gingers from Latest Cretaceous to Recent. *Science*, *289*(5477), 291-294.

Williams, C. A., Demissie, A. & Harborne, J. B. (1983). Flavonoids as taxonomic markers in old world *Lupinus* species. *Biochemical Systematics and Ecology*, *11*(3), 221-231.

Williams, W., Akhthar, M. A. & Faluyi, M. (1980). Cross compatibility between European and American *Lupinus* species. *Botanical Journal of the Linnean Society*, *81*(3), 225-233.

Wink, M. (1984a). Biochemistry and chemical ecology of lupin alkaloids. In *Proceedings of the Third International Lupin Conference* (p. 325-344). La Rochelle, France.

Wink, M. (1984b). Chemical defense of Leguminosae: Are quinolizidine alkaloids part of the antimicrobial defense system of lupins? *Zeitschrift für Naturforschung C*, *39*(6), 548-552.

Wink, M. (1984c). Chemical defense of lupins: Mollusc-repellent properties of quino-

lizidine alkaloids. *Zeitschrift für Naturforschung C*, *39*(6), 553-558.

Wink, M. (1992). The role of quinolizidine alkaloids in plant-insect interactions. In E. A. Bernays (Ed.), *Insect-plant interactions* (Vol. IV, p. 131-166). CRC Press, Boca Raton, Florida.

Wink, M., Meißner, C. & Witte, L. (1995). Patterns of quinolizidine alkaloids in 56 species of the genus *Lupinus*. *Phytochemistry*, *38*(1), 139-153.

Wojciechowski, M. F., Lavin, M. & Sanderson, M. J. (2004). A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. *American Journal of Botany*, *91*(11), 1846-1862.

Wolf, U., Ritter, H., Atkin, N. B. & Ohno, S. (1969). Polyploidization in the fish family Cyprinidae, order Cypriniformes. I. DNA-content and chromosome sets in various species of Cyprinidae. *Humangenetik*, *7*(3), 240-244.

Wolko, B. & Weeden, N. F. (1989). Estimation of *Lupinus* genome polyploidy on the basis of isozymic loci number. *Genetica Polonica*, *30*, 165-171.

Wolko, B. & Weeden, N. F. (1990a). Isozyme number as an indicator of phylogeny in lupins. *Genetica Polonica*, *31*(3-4), 179-187.

Wolko, B. & Weeden, N. F. (1990b). Relationships among lupin species as reflected by isozyme phenotype. *Genetica Polonica*, *31*(3-4), 189-197.

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B. & Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(33), 13875-13879.

Woodward, F. I. (1998). Do plants really need stomata? *Journal of Experimental Botany*, *49*(suppl. 1), 471-480.

Wright, S. I., Le, Q. H., Schoen, D. J. & Bureau, Th. E. (2001). Population dynamics of an *Ac*-like transposable element in self- and cross-pollinating *Arabidopsis*. *Genetics*, *158*(3), 1279-1288.

Wright, S. I., Ness, R. W., Foxe, J. P. & Barrett, S. C. H. (2008). Genomic consequences of outcrossing and selfing in plants. *International Journal of Plant Sciences*, *169*(1), 105-118.

Wu, C.-t. & Morris, J. R. (2001). Genes, genetics, and epigenetics: a correspondence. *Science*, *293*(5532), 1103-1105.

Wu, M., Li, L. & Sun, Z. (2007). Transposable element fragments in protein-coding regions and their contributions to human functional proteins. *Gene*, *401*(1-2), 165-171.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J. & Knaap, E. van der. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, *319*(5869), 1527-1530.

Xing, J., Wang, H., Belancio, V. P., Cordaux, R., Deininger, P. L. & Batzer, M. A. (2006). Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 17608-17613.

Yakusheva, A. S. & Svist, M. V. (2008). Population structure of lupin anthracnose in

Russia. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 443-446). Fremantle, Australia.

Yáñez-Ruiz, D. R., Martín-García, A. I., Weisbjerg, M. R., Hvelplund, T. & Molina-Alcaide, E. (2009). A comparison of different legume seeds as protein supplement to optimise the use of low quality forages by ruminants. *Archives of Animal Nutrition*, *63*(1), 39-55.

Yang, L., Jin, G., Zhao, X., Zheng, Y., Xu, Z. & Wu, W. (2007). PIP: a database of potential intron polymorphism markers. *Bioinformatics*, *23*(16), 2174-2177.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, *10*(6), 1396-1401.

Yang, Z. (2008). Empirical evaluation of a prior for Bayesian phylogenetic inference. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 4031-4039.

Yang, Z. & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, *15*(12), 496-503.

Yang, Z. & Rannala, B. (2005). Branch-length prior influences bayesian posterior probability of phylogeny. *Systematic Biology*, *54*(3), 455-470.

Yoon, H. S., Grant, J., Tekle, Y. I., Wu, M., Chaon, B. C., Cole, J. C. *et al.* (2008). Broadly sampled multigene trees of eukaryotes. *BMC Evolutionary Biology*, *8*(1), 14.

Young, N. D., Mudge, J. & Ellis, T. H. N. (2003). Legume genomes: more than peas in a pod. *Current Opinion in Plant Biology*, *6*(2), 199-204.

Yue, F., Shi, J. & Tang, J. (2009). Simultaneous phylogeny reconstruction and multiple sequence alignment. *BMC Bioinformatics*, *10*(Suppl. 1), S11.

Zamora-Natera, F., García-López, P., Ruiz-López, M., Ruiz Moreno, J., Pedrosa, M. & Muzquiz, M. (2008). Composition of alkaloids in seeds of *Lupinus mexicanus* (Fabaceae) and antifungal evaluation of the alkaloid extract. In J. A. Palta & J. B. Berger (Eds.), *Proceedings of the 12th International Lupin Conference — Lupins for health and wealth* (p. 216-219). Fremantle, Australia.

Zazula, G. D., Harington, C. R., Telka, A. M. & Brock, F. (2009). Radiocarbon dates reveal that *Lupinus arcticus* plants were grown from modern not Pleistocene seeds. *New Phytologist*, *182*(4), 788-792.

Zhang, J., Yu, C., Pulletikurti, V., Lamb, J., Danilova, T., Weber, D. F. *et al.* (2009). Alternative Ac/Ds transposition induces major chromosomal rearrangements in maize. *Genes & Development*, *23*(6), 755-765.

Zhang, X. & Wessler, S. R. (2004). Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(15), 5589-5594.

Zhang, X. H.-F. & Chasin, L. A. (2006). Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(36), 13427-13432.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G. K.-S. & Yu, J. (2006). KaKs_Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics*

Proteomics Bioinformatics, *4*(4), 259-263.

Zhang, Z., Ling, H. & Li, M. (2008). Mango: multiple alignment with N gapped oligos. *Journal of Bioinformatics and Computational Biology*, *6*(3), 521-541.

Zhu, H., Choi, H.-K., Cook, D. R. & Shoemaker, R. C. (2005). Bridging model and crop legumes through comparative genomics. *Plant Physiology*, *137*(4), 1189-1196.

Zuccolo, A., Sebastian, A., Talag, J., Yu, Y., Kim, H., Collura, K. *et al.* (2007). Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. *BMC Evolutionary Biology*, *7*(1), 152.

Zuckerkandl, E. & Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, *8*(2), 357-366.

Table des figures

Liste des tableaux

Vu par les co-directeurs de thèse,

Abdel-Kader Aïnouche (MdC) Marie-Thérèse Misset (PR)

Résumé

Les génomes sont des structures dynamiques, variant en taille et en composition, dans lesquelles les rétrotransposons jouent un rôle moteur. Dans ce cadre, nous nous sommes fixé trois objectifs de travail : 1) améliorer notre connaissance des relations phylogénétiques au sein du genre *Lupinus* (Fabaceae) par l'utilisation de nouveaux marqueurs nucléaires (ARNr-ETS et *SymRK*), 2) évaluer par amplification et par hybridation *in situ* la diversité, l'abondance et le rôle des rétrotransposons Ty1/*copia* et Ty3/*gypsy* dans les variations de taille de génome des lupins, et 3) séquencer, annoter et comparer une première région génomique disponible pour un lupin avec les régions homologues d'autres fabacées. La phylogénie obtenu améliore notre compréhension de l'histoire évolutive des lupins, et met en évidence des schémas de variation de taille de génome différents d'une lignée à l'autre. Les analyses de rétrotransposons révèlent que les éléments *copia* et *gypsy* contribuent de façon plus significative aux différences de taille de génome chez les lupins méditerranéens que chez les lupins africains et suggèrent différents modes et mécanismes d'évolution de la taille des génomes au sein du genre. À l'échelle locale (région du gène *SymRK*), nous confirmons la forte implication de ces éléments qui représentent 25 % de la région analysée chez *Lupinus angustifolius*.

Abstract

Genomes are dynamic structures, varying in size and composition, in which retrotransposons play a major role. In this context, our work aims at: 1) clarifying the phylogenetic relationships within genus *Lupinus* (Fabaceae) using additional nuclear markers (rRNA-ETS and *SymRK*), 2) assessing the diversity, abundance and role of the retrotransposons Ty1/*copia* and Ty3/*gypsy* in *Lupinus* genome size variation by amplification and *in situ* hybridization, and 3) sequencing, annotating and comparing a first genomic region available for lupine with homologous regions in other Fabaceae. The obtained phylogenetic framework improves our understanding of the evolutionary history of lupines, and when combined with the exploration of retrotransposons, highlights lineage-specific patterns of genome size variation. *Copia* and *gypsy* elements appear to contribute more significantly to genome size differences in Mediterranean lupines than in African lupines, suggesting different mechanisms involved in the genus. This was confirmed at the local scale (*SymRK* gene region) where these retroelements represent 25% of the analyzed region in *Lupinus angustifolius*.